Integrative analysis of transcriptional activity and genome architecture changes upon viral infections

Marco Alexander Michalski

The Babraham Institute

Girton College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2017

Integrative analysis of transcriptional activity and genome architecture changes upon viral infections

Marco Michalski

To study the interplay between spatial nuclear architecture and transcriptional activity during viral infections, I employed a genome-wide chromosome conformation capture approach (Hi-C) on infected murine and human cells and further enriched those libraries for genomic loci of interest and the viral genomes with biotinylated RNA baits. In parallel, I profiled newly transcribed RNA throughout the entire kinetic of murine cytomegalovirus (mCMV) infection in mice. Host genome rearrangement is a well-known phenomenon of mCMV infection but the underlying mechanisms are largely unknown. Furthermore, HPV infection can lead to cervical cancers in humans, with genomic instability and re-arrangements, leading to dysregulation of gene expression. Thus studying changes in genome architecture at early stages of HPV induced carcinogenesis can further our understanding on how certain integration events can provide a growth advantage.

In this study, I identified clusters of genes characterized by distinct kinetic profiles upon CMV infection in the mouse, which were associated with distinct functional terms. ATAC-Seq uncovered proximal promoter regions (PPR) that showed an over-representation of specific transcription factor binding sites in each of the clusters. These correlated well with the annotated functions of the associated clusters.

Further, I found that lytic mCMV infection is accompanied by local and global changes of chromosomal interactions in the host cell genome. Notably, chromatin properties, such as gene density, GC content and the association with the nuclear lamina, predict the structural dynamics upon infection and correlate well with transcriptional activity and changes thereof.

High-resolution interaction profiles for TSSs of highly induced or repressed genes, suggest that in general, enhancer-promoter interactions already form in untreated cells; and these pre-existing DNA-structures are not significantly altered but function through transient activation or repression of enhancers.

Finally, the viral genome showed a distinct pattern of open and closed chromatin late in infection. We found that the 7.2 kb viral intron displays the most open chromatin, and is highly enriched for chromosomal contacts with the host genome.

Hi-C and capture Hi-C revealed that both short- (~50 kb) and long-range (~1 Mb) interactions occur during the early stages of HPV induced carcinogenesis between the host and the integrated HPV16 genomes. Integration and direct interactions between the viral genome and the host DNA were shown to be associated with changes in host gene expression. In addition, insertion of the virus can disrupt normal host architecture.

In summary, this project pioneers the study of changes in nuclear architecture upon viral infection in man and mice. I uncover numerous structural features and changes of both the viral genomes and the infected host cellular genomes, and I demonstrate that these changes correlate with transcriptional activity.

Science knows no country, because knowledge belongs to humanity, and is the torch that illuminates the world.

Louis Pasteur

Declaration

This dissertation is the results of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text or the acknowledgement of assistance table.

This thesis is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

This thesis does not exceed 60,000 words, excluding bibliography, figures and appendices.

Cambridge, UK, September 2017

Marco Alexander Michalski

Summary

Host genome re-arrangement and transcriptional re-programming are well-known phenomena of murine cytomegalovirus (mCMV) infection but the underlying mechanisms are largely unknown. Furthermore, human papillomavirus (HPV) infection can lead to cervical cancers in humans, with genomic instability and re-arrangements, leading to dysregulation of gene expression. Thus studying changes in genome architecture at early stages of HPV induced carcinogenesis can further our understanding on how certain integration events can provide a growth advantage. To study the interplay between spatial nuclear architecture and transcriptional activity during viral infections, I employed a genome-wide chromosome conformation capture approach (Hi-C) on infected murine and human cells and further enriched those libraries for genomic loci of interest and the viral genomes with biotinylated RNA baits. In parallel, I profiled newly transcribed RNA throughout the entire kinetic of mCMV infection in mice.

I identified clusters of genes characterized by distinct kinetic profiles upon cytomegalovirus (CMV) infection in the mouse, which were associated with distinct functional terms. ATAC-Seq uncovered proximal promoter regions (PPR) that showed an over-representation of specific transcription factor binding sites in each of the clusters. These correlated well with the annotated functions of the associated clusters.

Further, I found that lytic mCMV infection is accompanied by local and global changes of chromosomal interactions in the host cell genome. Notably, chromatin properties, such as gene density, GC content and the association with the nuclear lamina, predict the structural dynamics upon infection and correlate well with transcriptional activity and changes thereof.

High-resolution interaction profiles for transcription start sites (TSSs) of highly induced or repressed genes, suggest that in general, enhancer-promoter interactions are already formed in untreated cells; and these pre-existing DNA-structures are not significantly altered but function through transient activation or repression of the interacting regulatory elements.

Finally, the viral genome showed a distinct pattern of open and closed chromatin late in infection. I found that the 7.2 kb viral intron displays the most open chromatin, and is highly enriched for chromosomal contacts with the host genome.

Hi-C and capture Hi-C revealed that both short- (~50 kb) and long-range (~1 Mb) interactions occur during the early stages of HPV induced carcinogenesis between the host and the integrated HPV genomes. Integration and direct interactions between the viral genome and the host DNA were shown to be associated with changes in host gene expression. In addition, insertion of the virus can disrupt normal host architecture.

In summary, this project pioneers the study of changes in nuclear architecture upon viral infection in man and mice. I uncover numerous structural features and changes of both the viral genomes and the infected host cellular genomes, and I demonstrate that these changes correlate with transcriptional activity.

Acknowledgements

My PhD has truly shaped who I am and taught me what a life in science involves. I owe a lot to many people for that privilege.

I would like to start by thanking my supervisor Peter Fraser for inviting me to his group and letting me pursue my own scientific interests. Your immense knowledge in and of the field certainly made my life as a PhD student much easier. I also want to thank my co-supervisor Lars Dölken, who initiated most of the work presented in this thesis. I am grateful for your guidance and your continuous support throughout the years, but foremost for letting me take on such interesting projects and trusting me to take them into new directions. Besides my supervisors, I would like to thank Mikhail Spivakov and Nick Coleman for their insightful comments and discussions.

The Fraser lab has been a tight knit group throughout my PhD, with many amazing people to thank– I would like to start by thanking Stefan Schönfelder for his rigorous lab training and the considerable time and effort he has put in throughout my entire PhD to help me bring this thesis together. My sincere thanks also goes to Jörg Morf, who made me go mad in the lab while discussing his endless crazy ideas, but who also became a good friend, and with the help of his little family actually helped to maintain my sanity. Notably there was some extreme speedreading of this thesis by both, Stefan and Jörg, for which I am extremely grateful. I thank past and current lab members of the Fraser lab for the stimulating discussions and for all the fun we have had in the last four years. Last and some might also say least, I would like to thank my companion Stephen “hamster boy” Bevan for all the lively discussions about science and sports (even though it is not a sport). You know nothing Stevie Beliebie!
I would like to thank the Babraham bioinformatics group for getting so good at hiding their despair at seeing my face pop around the corner, in particular Steven Wingett.

Without collaborations, science would not be the same and I had the privilege to work with some very talented and nice young scientists, who also became good friends. Emma, I will miss your “quick” questions, but I am happy that you found your dream job outside of academia. Jiao, your never-ending drive, in science and in life in general, is amazing and something I deeply admire. I wish the best of luck to both of you in the future!

I am hugely grateful and humbled by the friendship of my fellow PhD students including but not limited to Jack, Azad, Steph, Joanina, Tom, Lina and Michiel without whom I would have

lost my sanity long ago. You are all cheeky little devils and some of you are actually not that bad at climbing.

A big ‘Dankeschön!’ to my family and friends, who put up with me moving so far away for five years and only coming home occasionally. Your Skype calls and messages kept my spirits up. I would particularly thank my parents for opening up the world to me and supporting me wherever I go. No words can describe how grateful I am to have my sister. Even though I am clearly the smarter one, I have learnt a lot from her over the decades.

Finally yet importantly, I would like to express my gratitude for my girlfriend Jo. Not only are you the most cheerful person I know, you are also the most caring. I will be forever thankful for your support and your love. I could not have done that PhD without you.

Acknowledgement of assistance

1) Initial training in techniques and laboratory practice and subsequent mentoring:

Dr. Stefan Schönfelder – Training in Hi-C and capture Hi-C and other molecular biology techniques; mentoring and assessment

Dr. Jörg Morf, Dr. Mayra Furlan-Magaril – Training in molecular biology techniques

Dr. Peter Fraser, Dr. Mikhail Spivakov, Dr. Nick Coleman, Dr Lars Dölken – Mentoring and assessment

Dr. Steven Wingett, Dr. Csilla Várnai, Dr. Jonathan Cairns, Dr. Simon Andrews – Training in computational analysis

2) Data obtained from a technical service provider:

Kristina Tabbada (Babraham sequencing facility) – Bioanalyzer and Illumina sequencing of Hi-C, SCRiBL and gDNA libraries

Cambridge Genomic Services (CGS) – Illumina sequencing of ATAC-Seq libraries

BGI Hong Kong – Library generation and sequencing of newly transcribed RNA

3) Data produced jointly:

Dr. Emma Knight, Dr Ian Groves – Generation of Hi-C and SCRiBL libraries on W12 cell lines

Dr. Stefan Schönfelder, Stephen Bevan – Solution hybridisation of murine Hi-C libraries

Dr. Andrzej Rutkowski – purification of newly transcribed RNA

4) Data/materials provided by someone else:

Dr. Steven Wingett – mapping of gDNA-Seq

Dr. Csilla Várnai – PCA of mCMV infected cells, significant changes in compartmentalisation

Dr. Emma Knight – Fluorescent in situ hybridisation (FISH), PCR validation of HPV16 break points

Jack Monahan – Analysis of HPV16 Hi-C and SCRiBL data i.e. Circos plots, 4CSeq, heatmaps

Olga Mielczarek – Analysis of FISH experiments

Joana Guedes – Performance and analysis of confocal microscopy data

Dr. Anton Enright – Gene expression analysis of HPV 16 cell line

Dr. Florian Erhard – mapping of ATAC-Seq data

Dr. Caroline Friedel – mapping of 4sU-Seq data

Table of Contents

Abbreviations

3C	Chromosome Conformation Capture
3D	three dimensional
4sU	4-Thiouridine
AP	assembly protein
APOT	Amplification of Papillomavirus Oncogene Transcripts
BAC	bacterial artificial chromosome
bp	base pairs
BrUTP	bromouridine triphosphate
cDNA	complementary DNA
CFS	common fragile sites
CGI	CpG islands
ChIP	Chromatin Immunoprecipitation
CMV	cytomegalovirus
CNV	copy number variations
CT	chromosome territories
CTCF	CCCTC-Binding factor
CTD	c-terminal domain
DAPI	4',6-diamidino-2-phenylindole
DBD	DNA-binding domain
DBD	DNA-binding domain
gDNA	genomic DNA
DHS	DNase I Hypersensitive Sites
DMEM	Dulbecco's Modified Eagle's Medium
DMSO	dimethyl sulfoxide
DN-MT	DNA methyltransferases
dNTP	deoxynucleotide
DSB	double strand breaks
E	early
E6-AP	E6-associated protein
EBV	Epstein-Barr virus
EDTA	Ethylenediaminetetraacetic acid
EGF	epidermal growth factor
eRNA	enhancer RNA
FBS	Fetal bovine serum
FISH	fluorescent *in situ* hybridisation
GO	gene ontology
H3K27Ac	histone H3 acetylation at lysine 27
H3K27me3	histone H3 tri-methylation at lysine 27
H3K4me2	histone H3 di-methylation at lysine 4
H3K79me2	histone H3 di-methylation at lysine 79
H3K9me1	histone H3 mono-methylation at lysine 9
H3K9me3	histone H3 tri-methylation at lysine 9
HBV	Hepatitis B virus
hCMV	human cytomegalovirus
HCV	Hepatitis C virus
HDAC	histone deacetylases
HHV	human herpes virus

HIF-1	hypoxia inducible factor
HIV	human immunodeficiency virus
HP1	heterochromatin protein1
hpi	hours post infection
HPV	human papillomavirus
HRHPV	high-risk human papillomavirus
HSV	herpes simplex virus
IDT	Integrated DNA Technologies
IE	immediate early
IF	Interferon
IQR	interquartile range
kb	kilobases
kDa	kilo Dalton
KSHV	Kaposi's sarcoma-associated herpesvirus
L	late
LAD	lamina-associated domains
LCR	long control region
LSIL	low-grade squamous intraepithelial lesion
Mb	megabases
mc-BP	minor capsid binding protein
mCMV	murine cytomegalovirus
mCP	minor capsid protein
MCP	major capsid protein
MIEP	major immediate early promoter
mM	millimolar
MOI	multiplicity of infection
mRNA	messenger RNA
ncRNA	noncoding RNA
ND10	nuclear domain 10
NEMO	NF-κB essential modulator
NFR	nucleosome-free region
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	next generation sequencing
nt	nucleotides
ORF	open reading frame
Ori	origin of replication
p	passage
pA	polyadenylation
PBS	phosphate-buffered saline
PCA	principal component analysis
PcG	Polycomb group
PCR	polymerase chain reaction
PE	paired-end
Pen/Strep	Penicillin Streptomycin
PFU	Plaque-forming unit
PIC	pre-initiation complex
PML	promyelocytic leukemia protein
PPR	pattern recognition receptor

PPR	proximal promoter region
pRB	retinoblastoma protein
PRC	polycomb repressive complex
PRV	pseudorabies virus
P-TEFb	positive transcription elongation factor b
qRT-PCR	quantitative reverse transcriptase PCR
RNAP	RNA polymerase
RPKM	reads per million per kilobase of exon
rpm	revolutions per minute
rRNA	ribosomal RNAs
RT	room temperature
SCC	squamous cell carcinoma
SCP	small capsid protein
SCRiBL	sequence capture of regions interacting with bait loci
SDS	Sodium dodecyl sulphate
Shh	*Sonic Hedgehog*
SINE	Short Interspersed Nuclear Elements
TAD	topological Associated Domain
TBP	TATA-binding protein
TEM	transmission electron microscopy
TF	transcription factor
TFBS	transcription factor binding site
tRNA	transfer RNAs
TSS	transcription start sites
VRC	viral replication compartment

List of figures

List of tables

1 Introduction

Higher eukaryotes are made up of trillions of cells (3.72×10^{13} for the human body) (Bianconi et al., 2013) and are comprised of a large diversity of different specialised cell types, yet the majority of nuclei containing cells in an individual hold the same genetic information encoded in the DNA. The functions of a cell are mainly determined and performed by its protein repertoire, which is ultimately determined by the set of active and repressed genes in that cell (Lee & Young, 2013). Many cells must respond to environmental signals and stimuli, such as infections, through changes in gene expression (Marcinowski et al., 2012; Rutkowski et al., 2015).

One key level of gene expression regulation occurs at the level of transcription itself. A variety of different factors influence transcriptional regulation (Lelli et al., 2012), such as proximal and distal elements of DNA sequences, chromatin status and histone modifications at these regulatory sites, protein transcription factors (TFs) and other chromatin modifying proteins which can bind and interact with these elements and noncoding RNAs (ncRNAs). Gene transcription and silencing occurs within the three-dimensional (3D) context of the nucleus and thus the spatial organisation of genes and regulatory elements in the nucleus is emerging as an increasingly important aspect of gene regulation (Bonev & Cavalli, 2016).

In this thesis, I investigate the regulation of gene transcription following murine Cytomegalovirus (mCMV) infection in the context of 3D organisation of the host genome as well as the viral genome. I will further investigate the role of host genomic 3D architecture in early stages of human papillomavirus 16 (HPV16) induced carcinogenesis. In the following sections, I summarise the current knowledge of transcriptional regulation, particularly in the context of 3D organisation. I also recapitulate how viruses exploit the host cell machinery to facilitate their own needs.

1.1 Cytomegalovirus

Cytomegaloviruses (CMV) are herpes viruses (Herpesviridae) belonging to the order of Herpesvirales, which was newly recognized in 2009, encompasses about 120 different species and has birds, mammals and reptilians as hosts (Davison et al., 2009). Currently there are eight human pathogenic herpes viruses identified, responsible for a variety of diseases, such as common cold sores in the oral and genital area, chickenpox and proliferative diseases (Barozzi et al., 2007). The family is further divided into three subfamilies: Alphaherpesvirinae, Betaherpesvirinae and Gammaherpesvirinae, with CMV being part of the Betaherpesvirinae.

1.1.1 Associated diseases of human CMV (hCMV)

The human pathogenic hCMV is the most common infectious cause of congenital acquired damage in the central nervous system, such as hearing and visual impairments, and developmental delays (Cheeran et al., 2009; Damato & Winnen, 2002). Primary infection with this virus is usually asymptomatic. Transmission of the virus occurs via sexual contact, breast milk and urine. Blood transfusions and organ transplants are other potential sources of transmission (Pereira et al., 2007). The virus is found worldwide and depending on the standard of living, between 40 and 90 % of people are infected (Fulop et al., 2013). In the rare case in which primary infection becomes symptomatic, an infectious mononucleosis-like picture is observed (Klemola et al., 1970). Life-threatening complications in immune-competent adult humans are rare. In contrast, in immunocompromised patients, primary CMV infection or CMV reactivations commonly lead to serious complications if not recognized and treated. Typical presentations include CMV pneumonia, colitis or hepatitis in bone marrow or solid organ transplant patients. Following a kidney transplant, manifest CMV reactivation can lead to rapid deterioration of graft function and loss of the graft. In 30 % of all AIDS patients, who do not receive highly active antiretroviral therapy, CMV reactivates and spreads to and within the retina, commonly leading to blindness (Sittivarakul & Seepongphun, 2017). Primary infection of the mother during pregnancy poses a substantial risk for the unborn child. While the infection is usually hardly noted by the mother, the child often faces life-threatening complications. About 0.3-1 % of pregnant women get infected with the virus during pregnancy and in 40 % the infection is transmitted to the unborn child (Simonazzi et al., 2017). If the infection occurs in the first two thirds of pregnancy, it can cause deformities in the cardiovascular system, the gastrointestinal tract, the skeleton and the muscles. Furthermore, hepatosplenomegaly, microcephaly and chorioretinitis are observed. Consequently, two to three babies are affected by CMV viruses every day in the United Kingdom; almost 1,000 babies every year. Congenital CMV causes more birth defects and childhood deaths than Down’s syndrome, toxoplasmosis or listeriosis.

1.1.2 mCMV as a model for hCMV

CMVs are strictly host-specific, making it impossible to study hCMV in animal models. However, due to the similarity of the infection course, the establishment of latency and the similarity of diseases caused by mCMV and hCMV in their respective hosts, the infection of mice with mCMV is the most widely used model system to study hCMV (Reddehase et al., 2002; Sweet, 1999). Natural infection of the immunocompetent mouse proceeds as in humans and is usually not lethal with the virus establishing a symptom-free latent infection. In contrast, immunosuppressed mice respond to infection with fatal interstitial pneumonia, inflammation of the adrenal glands and hepatitis (Bolger et al., 1999; Brody & Craighead, 1974; Riddell, 1995).

Reactivation from latency is also comparable in mouse and man. At the gene level, both viruses share extensive homologies. 170 and 230 potential open reading frames (ORFs) were predicted by sequencing in the mCMV and hCMV genome, respectively (Chee et al., 1990; Rawlinson et al., 1996), 78 of which show significant amino acid homology to each other. Thirty-three of these 78 ORFs in turn can be found in all herpesviruses.

1.1.3 Assembly and genome of mCMV

The mCMV particle has the typical morphology of herpes viruses (Figure 1.1a). The virion consists of a typical 100 nm icosahedral large capsid that surrounds one single copy of the ~230 kb, linear double-stranded DNA genome. The tegument is encompassed by a lipid bilayer, the viral envelope. The mCMV capsid itself consists of seven proteins: the major capsid protein (MCP), the minor capsid protein (mCP), the minor capsid binding protein (mC-BP), the small capsid protein (SCP) and three different assembly proteins (AP1, AP2 and pp150) (Ye et al., 2017). The genes for these proteins are members of the conserved for all herpes viruses set of genes (Gibson, 1996). The tegument is located between the envelope and the capsid and is composed of numerous viral proteins. These do not only link the capsid to the virion membrane but also exert an important role in host cell modulation and disarmament of the innate immune response upon virus entry. The size of the matrix space increases the contrast between the envelope and the capsid in the transmission electron microscopy (TEM) image, causing the typical so-called "egg shape" of all Herpesviridae (Zhou et al., 1994). A representative TEM image of hCMV is depicted in Figure 1.1b.The first shell is formed in the nucleus, by virions emigrating through the inner nuclear membrane and is removed again by fusion with the outer nuclear membrane. The viral envelope is formed during the second envelopment of the capsid by budding from Golgi-derived vesicles and contains at least 10 viral glyco-proteins. These play an important role by mediating the interaction of extracellular virions with the corresponding cellular receptors (Mettenleiter, 2006).

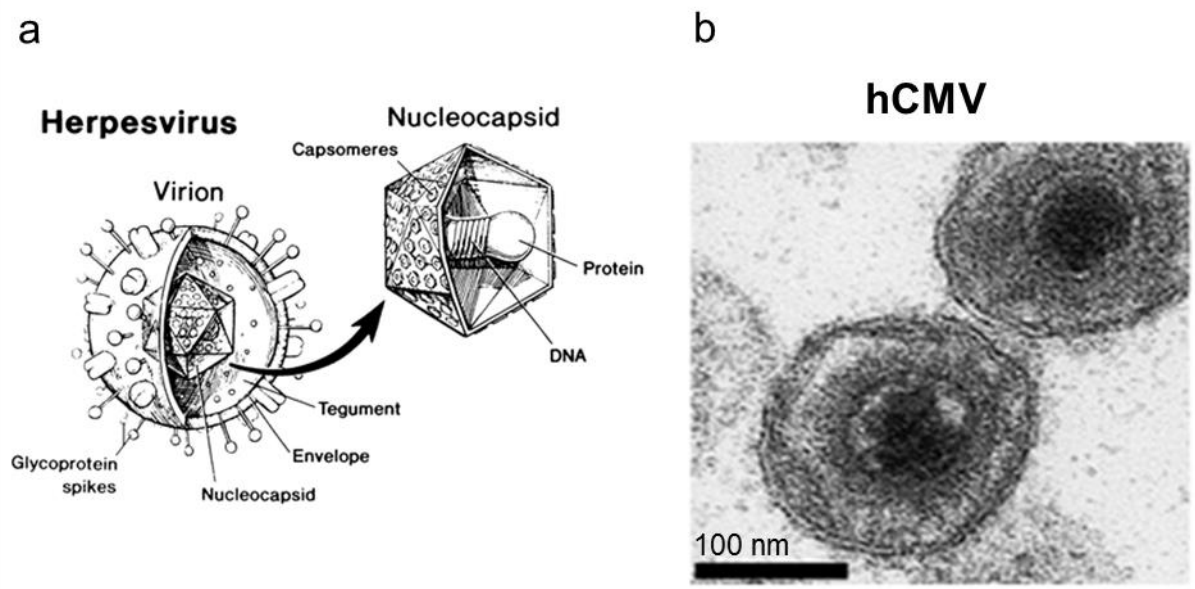


Figure 1.1 | The herpes virus particles

(a) A schematic model of a herpes virus particle and its main components; picture taken from (Liesegang, 1992) **(b)** Extracellular hCMV particle in a TEM image. Bar = 100 nm; picture taken from (Mettenleiter et al., 2009).

At a size of more than 230 kb, the CMV genome is the largest known herpes virus genome, and it is among the largest genomes of viruses infecting mammals. The mCMV genome is a single long sequence with short direct repeats at both ends (direct terminal repeats). Sequencing revealed a potential of 170 ORFs (Rawlinson et al., 1996), with genes residing on both strands and frequently overlapping, resulting in a very densely packed genome. Recent work employing ribosomal profiling suggests that these viruses might even encode a substantially larger number of gene products, i.e. > 700 for hCMV (Stern-Ginossar et al., 2012). In addition, the CMV genome has a conserved core region of 180 kb, which is orthologous to other herpes viruses. These genes are essential, and encode proteins that are involved in DNA replication, metabolism, and the assembly of the virion (Chee et al., 1990). Additionally, 18 miRNAs encoded by mCMV were identified (Dolken et al., 2007).

1.1.4 Life cycle of CMV

In general, CMV first binds to specific cell surface receptors and triggers the viral envelope to fuse with the cellular membrane. This leads to the release of capsids into the cytoplasm. Subsequently these capsids are transferred to the nuclear pore complexes where capsid disassembly (“uncoating”) occurs (Radtke et al., 2006). This transport is facilitated by cytoplasmic microtubules and dynein/dynactin motor proteins The nuclear import of the viral genome requires importin-β, and within 30 min post-infection the linear viral genome is converted into a covalently closed circular form (Boehmer & Nimonkar, 2003). In the nucleus, the cellular and viral protein machineries manufacture new viral DNA and proteins. After autocatalytic assembly of the capsids in the nucleus, newly synthesized viral genomes are

encapsidated. The nucleocapsid is then translocated to the cytoplasm by budding through the inner nuclear membrane. This process, together with the fusion of the primary envelope with the outer membrane and the release of the nucleocapsid into the cytoplasm, is called nuclear egress (Roller & Baines, 2017). In a next step, tegumentation and secondary envelopment take place. This includes a complex interaction of cytoplasmatic tegument proteins with the nucleocapsid and the future envelope resulting in infectious particles (Heming et al., 2017). As a cause of this lytic phase of infection, the cell is usually irreversibly destroyed.

A remarkable feature of all herpes viruses is their ability to establish a lifetime persistence in their hosts in the form of latency, after primary acute infection. During latency, the lytic transcription program is suppressed, no infectious virus particles are produced and viral transcription is restricted to the expression of a few latency-associated transcripts (Poole & Sinclair, 2015). Their task is to preserve the viral genome, to maintain latency and distract the host's immune system from the latently infected cell. During latency, the viral DNA remains in the nucleus in the form of an extrachromosomal episome. For some viruses, e.g. Epstein-Barr virus (EBV), this episome is replicated during cell division, coinciding with the replication of the host genome, by cellular DNA polymerase and passed on to daughter cells (Scott, 2017). Latency can be interrupted by extrinsic signals and conditions, such as immune cell depletion, allogenic transplantation and inflammatory disease resulting in reactivation of the virus, causing the biosynthesis and the release of new virus particles.

One crucial step for the establishment of latency and reactivation into the lytic phase is the regulation of the viral major immediate early promoter (MIEP). This promoter region controls the expression of the major viral immediate early (IE) genes: the cotransactivator IE1 for both mCMV and hCMV and the main early gene transactivator IE2 in hCMV and IE3 in mCMV, respectively. Since the activation of the IE genes is crucial for the activation of all other viral genes upon lytic infection, silencing of this promoter is a plausible mechanism to control viral lytic gene expression (Paulus & Nevels, 2009). Some cellular TFs like nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB), AP-1 and Sp1, have been shown to activate viral transcription from the MIEP, others, like YY1 and ERF, were demonstrated to repress its activation (Stinski & Isomura, 2008). The latter are involved in the recruitment of co-factors, which post-transcriptionally modify histones, thus mediating transcriptional repression (Weill et al., 2003; Wright et al., 2005). Chromatin Immunoprecipitation (ChIP) assays showed that the MIEP is associated with markers of repressed or active chromatin, according with its functional state (Ioudinkova et al., 2006; Murphy et al., 2002). The mechanisms involved in the establishment of latency and the induction of reactivation are, however, far from fully understood.

1.1.5 Gene expression cascade

Traditionally, CMV gene expression is categorized into immediate early, early and late genes, based on the time of transcription during infection (Wathen & Stinski, 1982). All viral transcripts are produced by cellular RNA polymerase II and are 5´capped and 3´polyadenylated. The vast majority of viral proteins are encoded by unspliced messages. One exception is the *IE* locus of mCMV. The *ie1-ie3* transcription unit is differentially spliced, giving rise to the corresponding proteins IE1 and IE3. Once the viral DNA has entered the nucleus, transcription of the immediate early genes is initiated. This requires no protein synthesis, because tegument proteins from the incoming particles act as viral TFs and can modulate IE gene expression (Liu & Stinski, 1992). As mentioned above, the MIEP controls the production of mRNAs encoding the viral major immediate early proteins (Lukac et al., 1994). The IE proteins then induce the expression of early genes, which are generally transcribed prior to DNA replication. Proteins translated by that time are mainly important to modulate the host cell to favor and to establish viral replication (Compton & Feire, 2007). Following the initiation of viral DNA replication, late gene expression occurs. These late genes can be further divided into the leaky-late genes, expressed at low levels in early infection and dramatically up-regulated at late times, and the "true" late genes, expressed exclusively after viral DNA replication. These genes are required for assembly and egress of the virions and the encoded proteins mediate the production and release of infectious virus particles (Tandon & Mocarski, 2012). The mechanisms restricting viral late gene expression to after DNA replication are still poorly understood. Interestingly, a study could demonstrate, using 4sU-tagging of nascent RNA, a sharp peak of viral gene expression during the first two hours of infection, including transcription of immediate-early, early and even well-defined late genes, at high and low multiplicity of infection (MOI). This is followed by rapid suppression of all three classes of viral gene expression by 5-6 hours post infection (hpi). Notably, this distinct peak for the early and late genes shifted to 3-4 hpi at low MOI. Further, it is important to note that the early burst of late transcripts at 1-2 hpi generates fully spliced and poly-adenylated mRNA (Marcinowski et al., 2012). Nevertheless, it remains unclear whether the respective "late gene" transcripts are actually translated. Another surprising finding of this study was the very constant rate of viral gene transcription or even continuously increasing suppression, despite the onset of extensive viral DNA replication.

1.1.6 Modulation of host gene expression

Compared to other herpes viruses, the infection cycle of CMV is rather slow, and no host shut-off takes place. Therefore, viral modulation of the cell has to be very effective. Like all herpesviruses, CMV have co-evolved with their hosts for millions of years. During this time, the virus has mastered host-cell modulation to facilitate its needs. Especially in the first hours of lytic CMV infection, various modulations of the host cell take place. Following attachment to

the cell, the virus is subjected to recognition by pattern-recognition receptors (PRRs) resulting in an innate immune response. This leads to activation of various cellular signal transduction pathways, including the NF-κB signalling pathway, and to the production of pro-inflammatory cytokines, such as the Interferon (IF) family (Isaacson et al., 2008). CMV has gained the ability to manipulate this early host inflammatory response towards its own needs. One important example of modulation of the host cell response by the virus is the interaction with the NF-κB pathway. First, CMV activates the NF-κB response resulting in favourable conditions for viral replication. However, later in infection virion-associated proteins as well as the onset of viral gene expression counteract the NF-κB response, thereby weakening the inflammatory host response (Browne et al., 2001; Montag et al., 2006). In this context, it was shown that the mCMV M45 protein plays a role in inhibiting the NF-κB response by proteasome-independent degradation of the NF-κB essential modulator (NEMO), hereby counteracting the host immune response (Fliss et al., 2012). More recently, the viral gene product M35 was identified to target NF-κB, but not IF-mediated transcription of the infected mouse cell, immediately upon infection (Chan & Goncalves Magalhaes, 2017).

Another example for overcoming the host immune response is the disruption of the nuclear domain 10 (ND10) bodies, named after their approximate average number per cell. One of the earliest events upon entry of the viral genome into the nucleus is the deposition to the ND10 bodies. This appears to be part of an intrinsic antiviral defence mechanism suppressing the expression of foreign DNA entering the nucleus, in part by recruiting chromatin-remodelling enzymes such as HDAC2 and transcriptional repressors such as Daxx (Tavalai & Stamminger, 2011). The formation of these bodies occurs *de novo* at sites of viral genomes, independent of viral transcription and very rapidly (within minutes). These aggregates contain various proteins such as HDAC2, Daxx, and promyelocytic leukemia protein (PML), which forms the matrix of the so-called PML-bodies (Chang et al., 2017). IF up-regulates the expression of those genes, and after infection some viral genomes are found beside or juxtaposed to ND10 bodies. During the first few hours of infection, the hCMV tegument protein pp71 is located at all ND10 bodies and ND10 bodies can be found juxtaposed to the hCMV major immediate early transactivator IE2, which co-localizes with TATA-binding protein (TBP) and TFIIB and is repressed by Daxx (Maul & Negorev, 2008). For hCMV, the tegument protein pp71 has been reported to destroy much of the cellular Daxx by a proteasome-dependent but ubiquitin-independent pathway (Hwang & Kalejta, 2007). No pp71 homolog has been identified in the mCMV genome and ie1 is the only mCMV protein known to interact with a ND10 body component, namely Daxx (Tang & Maul, 2003). Thus, the dispersion of ND10 bodies by mCMV seems to be predominantly mediated by ie1 (Martinez et al., 2010). Interestingly, a recent study could demonstrate for pseudorabies virus (PRV), a swine alphaherpes virus, that independent of the number of incoming genomes, only fewer than 7 genomes are actually expressed and

replicated (Kobiler et al., 2010). This finding correlates with the number of ND10 bodies per cell and might be one cause for the high ratio of virus particles to plaque-forming unit (PFU), observed for herpes viruses. It could be observed that genomes associated with ND10 bodies preferentially form viral replication compartments (VRCs) and that each replication compartment initiates from one genome (Kobiler et al., 2010). Whether this is also true for CMV is not known yet.

Like many other viruses, CMV dysregulates the host cell cycle machinery to its advantage. As such, CMV stimulates pro-proliferative cellular pathways, but once the infected cells reach the G1/S transition, the hCMV IE2 protein inhibits further cell cycle progression by specifically blocking cellular DNA synthesis (Wiebusch & Hagemeier, 1999). However, only DNA synthesis and cell division are blocked while other features of S-phase cells, such as an active nucleotide metabolism and the expression of replication factors, are still induced (Hertel et al., 2007). At the same time, cell cycle checkpoints are by-passed, preventing apoptosis (Gaspar & Shenk, 2006; Jault et al., 1995). The combination of this mechanism gives the virus sufficient time to replicate in a favourable environment. CMV also encodes proteins to target and inhibit the host cell's apoptosis response to prevent cell death (Brune et al., 2003; Goldmacher et al., 1999; Skaletskaya et al., 2001). These anti-apoptotic effects of CMV may permit prolonged viral infection and perhaps facilitate chronic inflammation.

During and after viral replication, the cell is reprogrammed to support production and release of infectious viral particles. However, viral modulations of the host at late times of infection are far less well studied and understood than the virus-host interactions during early phase of infection.

1.2 Human Papillomaviruses (HPV)

HPVs belong to the family of Papillomaviridae, comprising a diverse group of several hundred species of non-enveloped DNA viruses (de Villiers et al., 2004), infecting all mammals and other amniotes, such as birds, snakes and turtles. Over 200 HPVs subspecies have been characterised and are classified by genotype into five evolutionary groups based on the tissue they infect and their disease associations (Doorbar et al., 2012). Papillomaviruses are highly host- and tissue-specific. They replicate exclusively in cutaneous or mucosal epithelia (the basal layer of the body surface tissues) in particular areas of the body; HPV1, for example, tends to infect the sole of the feet, whereas HPV2 preferentially infects the palm of the hands, where they may cause warts. Infection, depending on the type, is usually asymptomatic or causes small benign tumours, known as papillomas or warts. However, papillomas caused by some types, especially HPV 16 and 18, carry the risk of becoming cancerous (Munoz et al., 2006). Mucosotropic HPVs are the most common sexually transmitted pathogens known to humankind.

1.2.1 HPV and cervical cancers

Viral infections are responsible for approximately 15 % of all cancer cases worldwide (zur Hausen, 1991). Malignancies include Burkitt lymphoma, Kaposi sarcoma and hepatocellular carcinomas, caused by infection with EBV, human herpes virus 8 (HHV-8/KSHV) and Hepatitis B and C viruses (HBV & HCV), respectively (Liao, 2006). Another example of a viral driven proliferative disease is cervical cancer, with almost 99.7 % of all cases caused by persistent infection and ineffective clearance of HPV. Therefore, HPV infection is accepted to be the major cause of cervical cancer (Durst et al., 1983; Walboomers et al., 1999). Furthermore, HPV infection has been attributed to the pathogenesis of other cancers including those of the vagina, penis, anus, vulva and oropharynx (Plummer et al., 2016), all of which are typically caused by the infection with one, or more, of the high-risk HPVs (HRHPV). These have higher oncogenic potential than low-risk HPV (LRHPV) types. HRHPVs including HPV16, 18, 31 and 45 are associated with over 90 % of cervical malignancies, with HPV16 alone accounting for over half of all cases worldwide (Scheurer et al., 2005; Zheng & Baker, 2006). Despite this, infection with HRHPV does not necessarily mean that cervical abnormalities will develop; only 0.3 % to 1.2 % of initial infections will eventually progress to invasive cervical cancer (Shulzhenko et al., 2014). Although HPV infection is the most common sexually transmitted disease, the overall prevalence of HRHPV infection is 23 % (Datta & Saraiya, 2011) and approximately 90 % of infections are spontaneously cleared by the host immune system within two years (Chen et al., 2014; Stanley, 2008). Persistent infection is seen in only 10-15 % of women who are unable to clear the infection; HPV persistence is the main risk factor associated with progression and, as a result, these women are at greater risk of developing cervical cancer (Pett & Coleman, 2007). In the case of cervical cancers, the HPV genome, which normally exists as a double stranded, circular and nuclear plasmid, is commonly found integrated into the host genome. It expresses two viral oncogenes, E6 and E7, which are implicated in the development and maintenance of the cancers.

1.2.2 HPV structure and genome

HPVs are relatively small, non-enveloped viruses of approximately 55 nm in diameter. They consist of an icosahedral capsid composed of 72 capsomeres containing the viral genome. Capsomeres are composed of two structural proteins: L1, which accounts for 80 % of the virus particle, and the minor capsid protein L2. The HPV16 genome is a double-stranded, circular (episomal) DNA molecule of 7904 bp. It is transcribed unidirectional and all resultant transcripts are polycistronic (Seedorf et al., 1985). The HPV16 genome is functionally divided into three regions: early, late and a long control region (LCR). These domains are separated by two polyadenylation (pA) sites. The early (E) region of the genome encodes six common ORFs: E6, E7, E1, E2, E4 and E5, which are required for the regulation of viral DNA replication and

viral gene expression. The late (L) region encodes for the structural proteins L1 and L2 that make up the viral capsid. The LCR region is an 850 bp non-coding, regulatory region that contains the origin of replication (Ori) as well as multiple *cis*-regulatory elements, including multiple transcription factor binding sites (TFBSs), such as NF1, AP1, Oct1 and TEF1. Binding of these TFs can synergistically modulate early promoter activity over a range of two to three orders of magnitude (Bernard, 2002). However, transcriptional activation can be antagonised through the binding of repressive TFs, for example, YY1 and CDP, which act as negative regulatory elements or silencers (Bernard, 2013). The HPV16 genome contains two major promoters. The p97 promoter lies upstream of the E6 ORF and is controlled primarily by upstream *cis*-elements in the LCR, which are responsible for early gene expression. The late promoter, p670, lies within the E7 ORF, is responsible for late gene expression and only is induced in differentiated keratinocytes (Zheng et al., 2006) (Figure 1.2).

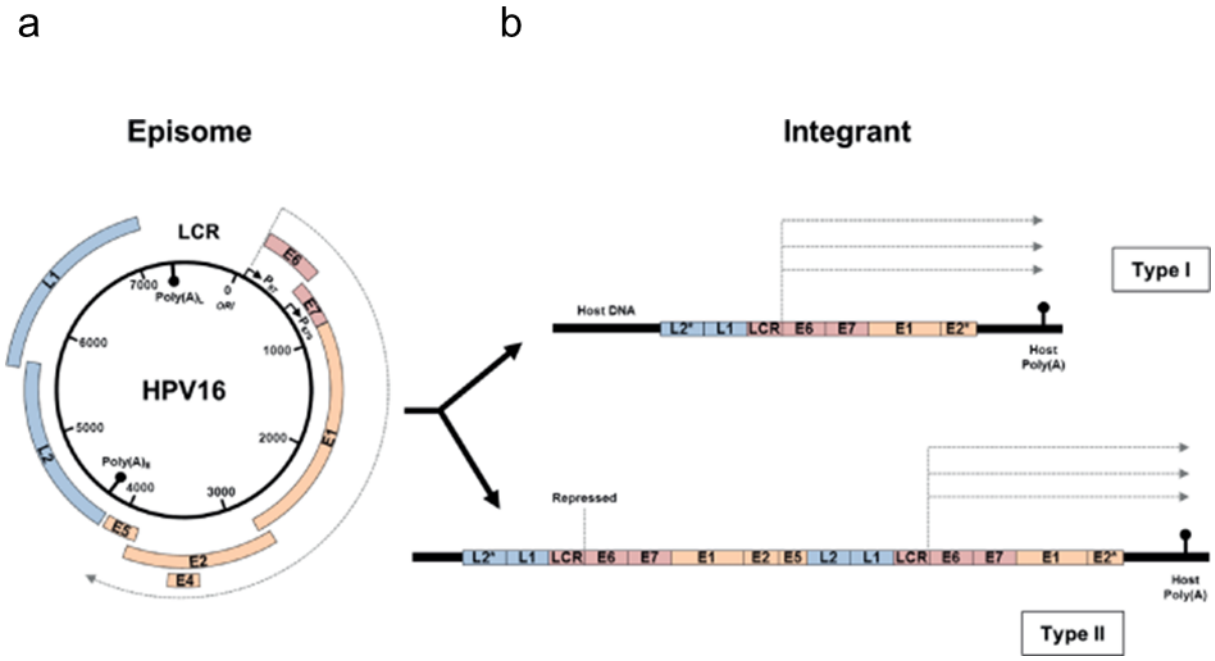


Figure 1.2 | Genome organisation and the physical state of the HPV genome

(a) The genomic organisation of HPV16 episome. The early (E), the late (L) and the long control region (LCR) are highlighted. The early (p97) and the late promoters (p670) and early (AE) and late (AL) polyadenylation sites are also indicated. **(b)** The two types of integrants of HPV observed in cervical neoplasia; type 1 retains E6/E7 but has a deletion or disruption of E2, whereas type 2 integrants have concatamerised full-length copies of the viral genome. Figure taken from (Groves & Coleman, 2015).

1.2.3 HPV16 oncogenes

In order to act as a carcinogenic agent, viruses are able to employ a variety of mechanism that result in cellular immortalisation and transformation. This occurs through the expression of viral oncogenes, which are able to inactivate regulators of genome stability, cell viability and cell cycle. The tumour suppressor proteins p53 and retinoblastoma protein (pRB) have been

shown to be targeted for degradation by a number of different viral oncogenes (Levine, 2009). HPV encodes for two viral oncogenes, E6 and E7, which function synergistically to enable limitless replicative potential, evasion of apoptosis, and genome instability, all of which are hallmarks of cancer (Hanahan & Weinberg, 2000).

The 150 amino acid (19 kDa) HPV E6 protein contains two zinc-binding regions, which enable the association with and degradation of numerous cellular proteins including the major cell-cycle checkpoint tumour suppressor protein p53 (Vliet-Gregg et al., 2013). E6 forms a complex with an E3-ubiquitin ligase called E6-associated protein (E6-AP), which is able to bind and ubiquitinate p53, leading to its proteasome-mediated degradation (Scheffner et al., 1993). Under normal conditions, activated p53 functions include the initiation of DNA repair pathways, cell cycle arrest, cell metabolism and/or apoptosis (Lane, 1992); however in the presence of E6, p53 cannot accumulate, and the ability of the cell to arrest cell cycle progression in response to DNA damage is removed. Additionally, E6 can enhance the degradation or the proteolytic inactivation of the pro-apoptotic proteins BAK and FADD, respectively (Yim & Park, 2005). This allows for the accumulation of mutations as cells with damaged DNA continue to replicate, rather than undergoing programmed cell death. Telomerase activation, leading to replicative immortality, is another proposed mechanism of HPV induced carcinogenesis (Yim et al., 2005). In most immortalised cells, including 85-90 % of cells derived from human cancers, the expression of telomerase is increased resulting in the maintenance of telomere length and the absence of cellular senescence (Hahn, 2002).

The HPV oncogenic protein E7 is an approximately 13 kDa phosphoprotein that contains a short CR2 motif, mediating interactions with and the degradation of pRB and its related proteins p130 and p107 (McLaughlin-Drubin & Munger, 2009), all of which are linked to cell-cycle control. This leads to E2F-regulated transcription, resulting in the production of cyclin A and cyclin E and ultimately inducing cell-cycle progression into S-phase and sustained proliferative signalling (Yim et al., 2005). Epigenetic reprogramming by modulating the activities of histone deacetylases (HDACs) and DNA methyltransferases (DN-MTs) is also caused by the viral E7 oncogene. E7 had been shown to stimulate DNMT1 *in vitro*; induction of this enzyme results in reduction of global levels of the repressive H3K27me3 histone mark and consequently the loss of polycomb repressive complex (PRC)-mediated repression (Munger & Jones, 2015). Additionally, E7 is able to bind and sequester HDACs, resulting in activation of several cellular promoters, including hypoxia inducible factor 1 (HIF-1), which causes increased levels of angiogenesis (Bodily et al., 2011). Finally, expression of E7 can also cause genomic instability by inducing centrosome amplification; this can lead to aneuploidy and structural chromosomal instability (Duensing & Munger, 2002). The overall result of E7 activity is to allow cell growth without differentiation, which can lead to immortalization.

The cooperative action of E6 and E7 leads to the emergence of a clonal population of cells with a growth advantage with a predisposition for transformation and malignant progression (Hickman et al., 2002; Moody & Laimins, 2010). This dogma has been illustrated using the W12 cell system (see 1.2.5).

1.2.4 HPV life cycle

Primitive basal keratinocytes are infected by HPV virions via microabrasions in the epithelial surface (Schiller et al., 2010; Stanley et al., 2007). The virus completely relies on the host cellular replication machinery, as the viral replication proteins E1 and E2 are insufficient to complete the replication of the viral genome (Conger et al., 1999). Once within the cell, the virus hijacks cellular resources to express proteins in a temporal and spatial pattern to facilitate replication of its own genetic material. Initially the viral genomes reside within the nucleus as extrachromosomal episomes and are passed on to daughter cells as the keratinocytes proliferate. At this stage of infection, viral genomes are maintained in low copy numbers (~10-200) through co-ordinated replication with the host DNA and early viral gene expression of particularly E1 and E2 (Doorbar, 2005). Upon differentiation of keratinocytes, the productive phase of the viral life cycle is initiated, resulting in induction of the early promoter and thus in increased levels of E6 and E7, which deregulate cell cycle control. The cell is now permissive to viral replication and viral copy numbers dramatically increase to thousands of copies within one single cell (Crosbie et al., 2013; Moody et al., 2010). Once the cell migrates further away from the basal layer and reaches the mid and upper layers of the epithelium, the viral late promoter (p670) becomes activated, resulting in the expression of the structural L1 and L2 capsid proteins, which encapsidate the newly synthesised viral genomes and produce infectious particles. The entire life cycle takes about 2-3 weeks, which correlates with the time necessary for cervical keratinocytes to undergo complete differentiation, migrate through the layers of the epithelium and desquamate (Crosbie et al., 2013; Stanley et al., 2007).

HPV infection results in very poor host humoral and cell-mediated immune response. Viral replication takes place in cells already destined to die as part of the normal process of skin shedding; as such, the productive cycle of HPV does not cause virus-induced cytolysis or necrosis, and thus does not lead to inflammation or the production of pro-inflammatory cytokines (Stanley et al., 2007). Furthermore, by avoiding immunocompetent cells at upper layers of the skin, the virus is able to establish a chronic infection, the single most important risk factor of cervical squamous cell carcinoma (SCC). Although the viral genome is maintained in an episomal form throughout the lytic phase, integration has been shown to correlate with the progression of precancerous lesions to invasive cancer (Pett et al., 2007; Wentzensen et al., 2004); indeed integration has been identified in 86.5 % of all cervical SCCs (Hu et al., 2015). For the viral genome to integrate, it commonly breaks in the E1 and/or E2 ORF, resulting

in loss of those genes (Zhao et al., 2016) and the loss of the E2-dependent negative feedback of early promoter expression. In contrast, the integrated genomes of selected cells faithfully retain the LCR and full length E6 and E7 ORF, which are both highly expressed. As a result, cell proliferation, immortalisation and genomic instability are induced. Integration is not part of the normal life cycle of HPV, and HPVs encode no integrases or polymerases. HPV integration presumably occurs following double strand breaks (DSBs) in host and viral DNA, which may explain the frequency of integrations occurring at common fragile sites (CFSs) (Winder et al., 2007).

1.2.5 W12 cell lines

Longitudinal investigations of cervical neoplastic progression *in vivo* are difficult to perform as, once detected, the disease is treated immediately. However, *in vitro* models are not subject to the same constraints and allow unique insight into the development of the invasive phenotype. The W12 model is a unique example of such a system. The “parental” W12 cell line is a polyclonal population of cervical squamous cells that were generated by explant culture of a naturally occurring HPV16-positive cervical low-grade squamous intraepithelial lesion (LSIL) (Stanley et al., 1989). At early passages, the HPV16 genome is maintained at ~100-200 episomal copies per cell, and when grown in organotypic culture recapitulates an LSIL phenotype (Pett et al., 2004). Individual culture series have been established by independent long-term *in vitro* cultivation. Upon continuous passage over 9-12 months, HPV-infected W12 cells mirror the virus and host events seen in neoplastic progression *in vivo*. The most frequent outcome is the breakdown of episomal persistence, with emergence of cells containing 1-10 copies of integrated HPV16. Integration of HPV16 causes transcriptional deregulation of the virus, resulting in an increased level of the oncoproteins E6 and E7, as well as genomic instability and can result in progression to carcinomas. Hence, the W12 cell system is widely recognised as the best available model for HPV16 driven cervical carcinogenesis.

In a previous study, limiting dilution cloning from an early passage of polyclonal parental W12 cells was performed under non-competitive conditions. This generated a panel of twenty-four clones that all arose from a common genetic background and differ only by the site of HPV16 genome integration into host chromosomes (Dall et al., 2008). As such, the range of integration events that exist prior to episome clearance and integrant emergence were identified, regardless of whether they had a selective advantage in mixed cell populations. Hence, these W12 clones (Dall et al., 2008) represent a unique system to examine the host and virus factors that determine the selection of a particular HPV16 integrant from the range that exists in a typical polyclonal population of pre-malignant cervical keratinocytes.

1.3 Nuclear organization

The eukaryotic cell nucleus is a very heterogeneous organelle, which is highly structured and non-randomly organized. The conformation of the genome is adapted to cell-, tissue- and differentiation-specific transcriptional programs; however, the relationship between the 3D organization of the genome and transcription is not fully understood (Bonev et al., 2016; Krijger & de Laat, 2016).

More than a century ago, researchers started making inquiries into the organization of the nucleus, the largest and most easily discernible organelle in the eukaryotic cell. Early in the 20th century, the first subnuclear structures were identified and later named Cajal bodies after their discoverer Santiago Ramon y Cajal (Gall, 2003). Twenty-five years later, Emil Heitz observed differentially staining of chromatin in interphase nuclei of mosses and described it as heterochromatin and euchromatin (Heitz, 1928). The realization that the nucleus contains genetic material in form of DNA fibres further fuelled the interest in the nuclear structure. In humans, the genome consists of more than 3 billion nucleotides and is contained in 22 pairs of autosomes and two sex chromosomes. When unwound and aligned end to end, the DNA measures roughly 2 m per nucleus, this is about 200,000 times the diameter of an average mammalian cell nucleus (de Wit & de Laat, 2012). Yet the genome is not only contained within a sphere smaller than one tenth of the thickness of a human hair (10 micron), but also functions within that space. Therefore packing DNA inside the nucleus imposes tremendous organizational challenges and suggests that the genome cannot exist as a simple one-dimensional polymer. While already conceptually interesting, the shape of the genome becomes even more fascinating when one realizes that it also relates to genome function at the gene level as well at the global nuclear level.

1.3.1 The linear genome sequence

The advent of next generation sequencing has made the full genomic sequences of many organisms available. It is clear that only a small fraction of the higher eukaryotic genome is encoding for proteins (Lander et al., 2001; Waterston et al., 2002). For example, less than 1.5 % of the human and mouse genomes are protein-coding sequences. It also became apparent that the number of protein coding genes in different organisms does not correlate well with biological complexity (Taft et al., 2007). Humans and mice possess a very similar number of protein coding genes (21,976 and 22,707, respectively), but the genome of the relatively simpler nematode worm *Caenorhabditis elegans* holds a comparable number of protein coding genes (20,517) (www.ensemble.org). This has led to the perceived paradox in how increasing biological complexity is programmed in higher eukaryotic organisms (Hahn & Wray, 2002; Taft et al., 2007). The increased biological complexity of higher eukaryotes appears to be largely due to an increase in the complexity of regulatory networks. In prokaryotic

systems the number or regulatory proteins appears to scale approximately quadratically with genome size, which is consistent with theoretical models (Mattick & Gagen, 2005). Therefore, regulatory information is likely to be the major content of information of the genome in complex organisms (Gagen & Mattick, 2005). It has been shown that prokaryotes have hit the evolutionary limit, intrinsic to primarily using protein based regulators due to the accelerating cost of regulatory architecture (Gagen et al., 2005). This, in combination with the observation that the protein repertoire is very similar between organisms of different complexities, implies that the additional regulatory information must reside within the non-coding sequences of the genome (Amaral & Mattick, 2008; Levine & Tjian, 2003; Taft et al., 2007). In fact, the amount of non-coding DNA in eukaryotic genomes is the only factor which appears to correlate well with biological complexity (Taft et al., 2007). Genomic regulatory information can reside in non-coding DNA at promoters adjacent to transcription start sites (TSS), but, especially important in more complex organisms, can also reside in distal intronic and intergenic regions (Bulger & Groudine, 2011; Lelli et al., 2012; Levine et al., 2003). These elements can recruit a particular set of TFs, or be modified to particular chromatin states in order to provide regulatory input to control transcription of specific genes.

1.3.2 The eukaryotic promoter

Promoters are regions of DNA, immediately upstream to TSS, which initiate transcription of genes and are the proximal point at which multiple regulatory inputs converge to control transcription initiation by the recruitment and positioning of the pre-initiation complex (PIC). Mammalian promoters can be divided into two categories, conserved TATA box–enriched promoters, which initiate at a well-defined “sharp” TSS, and more plastic, “broad” CpG-rich promoters (Carninci et al., 2006). The majority of TATA box containing promoters mediate tissue specific expression in adult cell types, whereas “broad” TATA box lacking promoters are instead enriched for CpG islands (CGIs, sequences enriched for CG dinucleotides compared to genome wide levels) and in many cases are ubiquitously expressed and are associated with house-keeping functions (Carninci et al., 2006; Forrest et al., 2014; Lenhard et al., 2012). However, some CGI-promoters are regulated in a cell-type specific manner through Polycomb recruitment and are associated with developmental genes.

In order to facilitate compaction and to provide another layer of regulation, DNA is not naked within the nucleus, but is associated with histones to form chromatin. The basic structural unit of chromatin is called nucleosome, which is formed from the wrapping of approximately 147 bp DNA around one histone octamer consisting of two copies of each core histone: H2A, H2B, H3 and H4 (Luger et al., 1997), although additional histone variants exist. The linker histone H1 interacts with the spacer DNA (30-50 bp) at the edge of the nucleosome core and facilitates tight packaging of chromatin. Further, nucleosomes are able to serve as docking platforms for

many nuclear proteins that modify the structure of chromatin, and hence influence chromatin organisation.

Genome-wide studies of nucleosome positioning, utilising MNase-Seq (Barski et al., 2007) and more recently ATAC-Seq (assay for transposase accessible chromatin followed by sequencing) (Buenrostro et al., 2013), have shown that there is a preferential organisation of nucleosomes at gene promoters and TSSs, especially at the promoters of active genes. DNA sequence preferences can influence nucleosome positioning, but also various ATP-dependent chromatin-remodelling complexes, such as SWI/SNF, can actively reposition or evict nucleosome cores (Jiang & Pugh, 2009). Very frequently, promoters contain a nucleosome-free region (NFR) just upstream of the TSS, which allows binding and initiation of the transcription machinery. The -1 positioned nucleosome can regulate accessibility of proximal regulatory elements. Of all the nucleosomes found in and around genes, the +1 nucleosome displays the tightest positioning preference and often contains histone variants H2A.Z and H3.3, facilitating histone eviction and PIC assembly. Even though increased accessibility is a general feature of active promoters, recent work has shown that CGI-enriched ("broad") developmental promoters can be accessible even when silent. In contrast, TATA-enriched ("sharp") promoters can have a nucleosome at the TSS when active, but show less well-defined positioning of other nucleosomes (Forrest et al., 2014; Jiang et al., 2009; Rach et al., 2011).

1.3.3 RNA transcription in eukaryotes

There are three different nuclear RNA polymerase complexes in eukaryotic cells, each of which recognises different promoters and transcribes different classes of genes. Ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and a few other ncRNAs are transcribed by RNA polymerases I and III, most of which are constitutively expressed (Dieci et al., 2007; Russell & Zomerdijk, 2006). RNA polymerase II (RNAPII) transcribes most messenger RNAs (mRNAs) from protein coding genes, and therefore is subject to much greater regulatory input compared to the other two RNAPs (Kornberg, 1999).

RNAPII requires multiple general TFs for site-specific transcription initiation, and in fact cannot even bind DNA on its own. It interacts with DNA as a large holoenzyme complex to form a PIC at core promoter regions (Lee & Young, 2000; Liu et al., 2013). The minimum set of general TFs is composed of TFIIB, D, E, F and H. The recruitment and assembly of the PIC at promoters is one of the major levels at which transcription is regulated. The Mediator co-activator complex is involved in bridging between other factors and the P (Borggrefe & Yue, 2011). Mediator binds unphosphorylated polymerase and delivers it to the promoter. Once incorporated into the PIC, it strongly stimulates the c-terminal domain (CTD) kinase of basal transcription factor TFIIH to phosphorylate the CTD of RNAPII at serin 5 to form Ser5P. This

phosphorylation disrupts mediator binding and releases it from the PIC (Buratowski, 2009). This Ser5P form of RNAPII is predominantly found around promoters, and binds the capping enzyme, which adds a 5`-methylated guanosine triphosphate cap to the RNAPII products. Once the recruited RNA polymerase II molecules initiate transcription through melting the DNA, they generally transcribe a short distance, typically 20–50 bp, and then pause. This process is controlled by the pause control factors DSIF and NELF, which are physically associated with the paused RNA polymerase II molecules (Yamaguchi et al., 2013). The paused polymerases may transition to active elongation through pause release, or they may ultimately terminate transcription. The PIC initially undergoes several rounds of abortive initiation, synthesising small RNA species (Lee et al., 2013). Pause release and subsequent elongation occur through recruitment and activation of positive transcription elongation factor b (P-TEFb), which phosphorylates the paused polymerase at the CTD on Ser2 forming Ser2P, whilst Ser5P is removed by a phosphatase (Buratowski, 2009). This form of phosphorylated RNAPII is generally associated with elongation, and mediates interactions with other proteins such as transcription terminators and RNA processing enzymes (Buratowski, 2009). Following initiation, RNAPII alone is capable of RNA transcript elongation and proofreading. In addition to the 5`-cap, RNA molecules produced by RNAPII are further modified by splicing and usually undergo 3` cleavage and polyadenylation (Lee et al., 2000).

1.3.4 Chromatin modifications

Histones are globular proteins with a flexible N-terminus (widely considered to be the tail) that protrudes from the nucleosome. Various enzymes can post-translationally modify these tails and specific chromatin modifications are enriched around TSSs, particularly at the +1 nucleosome. Many of the histone tail modifications correlate very well with chromatin structure and both, histone modification state and chromatin structure, correlate well with gene expression levels (Bannister & Kouzarides, 2011). Traditionally, chromatin was partitioned into two states based on its compaction in interphase, densely packed and stained heterochromatin and lighter stained, more open euchromatin, as detected by light microscopy (Heitz, 1928). The compaction of chromatin is a major determinant of the transcriptional activity, with heterochromatin typically representing a repressive state (Brown, 1966) and euchromatin the site of transcriptional activity (Chesterton et al., 1974).

1.3.4.1 Heterochromatin

A major function of heterochromatin is to protect the underlying DNA from being accessed by dedicated machineries and, thus, used for transcription or for other DNA-based transactions, such as repair. There are two recognized types of heterochromatin, which are separated by their distinct regulatory roles and the protein machineries they recruit, although the consequence is chromatin compaction in both cases. On one hand, *constitutive*

heterochromatin is typically hypoacetylated and marked by histone H3 tri-methylation at lysine 9 (H3K9me3), which provides a binding side for heterochromatin protein1 (HP1) (Bannister et al., 2001; Lachner et al., 2001). The majority of constitutive heterochromatin is found at the gene-poor pericentromeric regions of chromosomes, but it is also found at the telomeres and distinct regions dispersed throughout eukaryotic chromosomes. This type of chromatin most commonly resides in the periphery of the nucleus attached to the nuclear membrane and is more or less consistent between cell types within a species (Saksouk et al., 2015). On the other hand, *facultative* heterochromatin is rich in tri-methylation of H3K27 (H3K27me3), a mark put on by Polycomb group (PcG) proteins, which are regulating single-copy genes in a cell type- and developmental stage-specific manner (Sparmann & van Lohuizen, 2006). This type of chromatin is also most frequently found at the nuclear periphery. Comparison of genome-wide binding maps of HP1 (constitutive) and PcG (facultative) indicate that the two heterochromatic states are non-overlapping (de Wit et al., 2007; Filion et al., 2010). Additionally, these two chromatin states are distinct from each other in super-resolution imaging data, which reveal distinct chromatin packaging for the different epigenetic states, including active, inactive and PcG repressed domains (Boettiger et al., 2016).

Artificially targeting HP1 to active chromatin, often results in gene silencing and spreading of heterochromatin (Li et al., 2003; Verschure et al., 2005), and repositioning a locus to heterochromatic regions results in gene silencing (Csink & Henikoff, 1996; Dernburg et al., 1996; Harmon & Sedat, 2005). Conversely, positioning away from heterochromatic regions is associated with active histone modifications and transcriptional activation (Hendzel et al., 1998; Schubeler et al., 1997). These observations do not always hold up though, as many genes bound by HP1 are transcribed (Vakoc et al., 2005) and transcriptional activity has been observed within heterochromatic regions at the nuclear periphery (Luo et al., 2009).

1.3.4.2 Euchromatin

In contrast to heterochromatin, euchromatin is viewed as being the predominant form of transcriptionally active chromatin (Zhou et al., 2011). Many histone modifications including acetylation, methylation, phosphorylation and ubiquitination have been implicated in active transcription (Berger, 2007) and are enriched around TSSs, particularly at the +1 nucleosome. For example, H3K4me3 is a prominent mark at active promoters (Bannister et al., 2011) and in mammals is deposited by COMPASS-like complexes containing one of the four MLL proteins or SET1A/B (Shilatifard, 2008). H3K4me3 has been shown to interact with the initiation factor TFIID and thus stimulating PIC formation (Lauberth et al., 2013). This modification can also recruit, either directly or indirectly, other chromatin modifying enzymes, such as histone acetylases and deacteylases, histone demethylases and chromatin remodelling complexes (Bannister et al., 2011; Huang et al., 2006; Shi et al., 2006; Sims et al., 2007). Other

modifications frequently found at active promoters include multiple acetylation modifications of H3 and H4, including H3K9/14Ac (Liang et al., 2004), as well as the methylation marks H3K4me2 and H3K79me2 (Schubeler et al., 2004). Acetylation has the effect of changing the overall charge of the histone tail from positive to neutral, which disrupts electrostatic interactions between histones and DNA, resulting in a less compact chromatin structure. This can enable binding of TFs and the PIC (Wakamori et al., 2015; Wang & Hayes, 2008). One mark associated with active transcriptional elongation is H3K36me3, which is commonly found at gene bodies of transcribed genes, especially towards the 3` end, which makes it the only covalent histone modification enriched towards the end of genes (Bannister et al., 2005; Mikkelsen et al., 2007). This modification is thought to prevent inappropriate initiation within genes (Carrozza et al., 2005) and can be targeted by the CTD phosphorylation state of elongating RNAPII (Buratowski, 2009).

1.3.4.3 Boundary elements

The distance of heterochromatin spreading is stochastic, and therefore the formation of boundaries, which block the spreading of heterochromatin, is critical for maintaining stable gene expression patterns. In most cases, the disruption of normal gene expression patterns severely compromises an organism's fitness or health, as seen in a number of human diseases linked to uncontrolled heterochromatin spreading (Kleinjan & Lettice, 2008). In general, heterochromatin regions are flanked by DNA sequences termed boundary elements, which form fixed borders accompanied by sharp transitions in histone modification profiles. Such elements control the precise determination of epigenetic states, even when heterochromatin protein levels change. In other cases, borders are determined by the local balance of heterochromatin and euchromatin proteins, which tends to differ between cells. Such boundaries are termed negotiable borders (Kimura & Horikoshi, 2004). For example, in budding yeast, the balance between histone acetyltransferase Sas2-mediated acetylation of H4K16 and Sir2-mediated deacetylation of the same residue defines the borders of heterochromatin at telomeric regions (Kimura et al., 2002). In most cases, specific DNA elements demarcate the borders of heterochromatin regions and function as boundaries to prevent spreading of heterochromatin. These boundary elements are frequently bound by specific factors, such as CCCTC-Binding Factor (CTCF), that play a role in maintaining the boundary between distinct chromatin types (Ong & Corces, 2014). Boundary elements are enriched for certain modifications, such as H3K9me1, and are devoid of others, such as histone acetylation (Barski et al., 2007). Furthermore, a specific histone variant, H2A.Z, is highly enriched at these sites (Barski et al., 2007). Since heterochromatin spreading depends on cycles of histone modifications of adjacent nucleosomes, it is not surprising that certain DNA sequences that are known to exclude nucleosome assembly can also efficiently establish

heterochromatin boundaries (Bi et al., 2004). How all of these factors work together in order to maintain these boundaries is far from clear, but their importance is undeniable.

1.3.4.4 Chromatin states

Traditionally, chromatin has been divided into heterochromatin and euchromatin. There is now ample evidence that a finer classification is required. Integrative analysis of genome-wide binding maps of 53 chromatin components in *Drosophila* cells has shown that the genome is segmented into five principal chromatin types, which are defined by unique yet overlapping combinations of proteins. These chromatin states not only differ in their protein composition, but also in their biochemical properties, transcriptional activity, histone modifications, replication timing, and ability to associate with DNA binding factors (Filion et al., 2010). The authors discovered, in addition to the known HP1 and PcG heterochromatin, a novel type of repressed chromatin, which accounts for roughly half of the genome. This new type of chromatin is particularly gene poor, late replicating and closely associated with the nuclear lamina, but lacks binding of both stereotypical heterochromatin proteins. Furthermore, the authors found that transcriptionally active euchromatin consists of two types that differ in molecular organisation and H3K36 methylation and regulate distinct classes of genes. A more recent study took this approach even further and sub-divided the *Drosophila* genome into 30 chromatin states based on 73 chromatin modules, each state with distinct functional properties (Zhou & Troyanskaya, 2016). Together, these findings suggest that the classical segregation of chromatin into either transcriptionally inert heterochromatin or transcriptionally active euchromatin is too simplistic and finer levels of genome organisation influence the transcriptional status of a gene.

1.3.5 Regulation of transcription by distal sequence elements

Gene expression is regulated through the integrated action of many *cis*-regulatory elements, including core promoters and promoter-proximal elements as well as a various *cis*-regulatory modules that are localized at greater distances from the TSSs on the linear sequence, such as enhancers, silencers, insulators and tethering elements (Spitz & Furlong, 2012). It is becoming increasingly apparent that in multicellular organisms, the precise spatiotemporal expression patterns of tissue-specific genes is primarily established by distal regulatory regions (Bulger et al., 2011; de Laat & Duboule, 2013; Levine et al., 2003). Both sharp and broad tissue specific promoters can be regulated by distal elements, but it is particularly important for developmentally vital broad promoters, which tend to be controlled by larger numbers of elements located at greater linear distances (Engstrom et al., 2007; Lenhard et al., 2012; Ong & Corces, 2011). Enhancers are DNA sequences which contain multiple binding sites for sequence specific TFs (Spitz et al., 2012) and activate transcription at target promoters (Shlyueva et al., 2014). These sequences are seemingly able to act independently

of their genomic location or orientation relative to their target genes (Banerji et al., 1981) and are thought to physically interact with promoter via looping (de Laat et al., 2013). Silencers behave in a similar fashion but harbour binding sites for repressing proteins and thus negatively regulate gene expression of the target gene (Brand et al., 1985; Kolovos et al., 2012). Insulators are sequences which, in combination with bound protein factors such as CTCF, are able to prevent regulatory interactions between enhancers/silencers and promoters or prevent spread of heterochromatin (Gaszner & Felsenfeld, 2006).

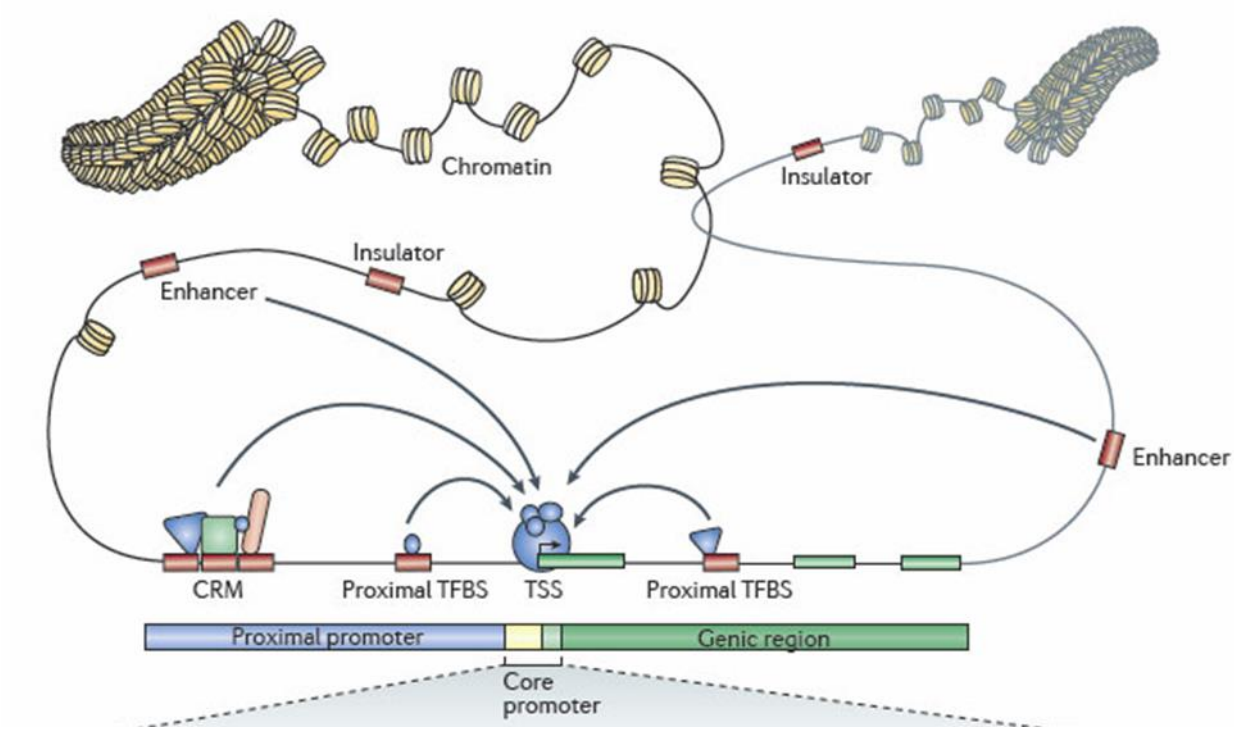


Figure 1.3 | Chromatin landscape of gene promoters

The TSS is a hotspot of transcriptional regulation. TFBS often appear in TSS proximal regions up- and downstream of gene promoters. TFBS often appear in clusters forming cis-regulatory modules (CRM). Along with the binding motifs present in the core-promoter (such as TATA-box), these CRMs anchor the binding sites for regulatory proteins that are able to recruit RNAPII to initiate transcription. Enhancer and silencer elements also bind regulatory proteins that can initiate or block transcription through spatial looping interactions. Insulator proteins mark boundaries between chromatin domains, and thus hinder looping interactions. Source: (Lenhard et al., 2012)

In the following section, I will give an overview of extensively studied enhancer elements to illustrate the characteristics and their regulatory function.

The human and mouse β-globin gene locus is possibly the most extensively studied region for spatial regulation (Yun et al., 2014). In erythroid cells where the gene is active, a specific loop is formed between its promoter and the upstream Locus Control Region, which consists primarily of DNase I Hypersensitive Sites (DHSs). Moreover, during red blood cell maturation, the gene relocates from the nuclear periphery to a transcription factory in the nuclear interior (Ragoczy et al., 2006). Cell type specific TFs TAL1 or KLF1 have also been indicated as key

factors in the unusual high expression rate of globin genes in erythroid cells (Schoenfelder et al., 2010b). Looping interactions between gene promoter and enhancers or silencers have been studied in great detail for several important genes. Some of these interactions span over very long distances in the linear genomic sequence: the *Myc* oncogene has an important enhancer ~330 kb away from the gene promoter (Pomerantz et al., 2009), and *Sonic Hedgehog* (*Shh*) has a limb-bud specific enhancer in a distance over a megabase (Amano et al., 2009).

Another remarkable feature of enhancers is that they are transcribed into RNAs (eRNAs) that do not encode for proteins and run the length of the enhancer sequence. Furthermore, eRNA production appears to serve as a more robust indicator of enhancer activities than any enhancer-bound transcription activators or chromatin marks (Raab & Kamakaka, 2010; Wang et al., 2011; Xu & Ling, 2017).

1.3.6 Transcription factor binding

TF proteins mediate a key aspect of transcriptional regulation at the above-mentioned gene regulatory regions. A defining feature of TFs is that they contain at least one DNA-binding domain (DBD), which binds a specific sequence of DNA proximal to the genes that they regulate (Brent & Ptashne, 1985). Local DNA accessibility of these motifs, defined by nucleosome occupancy, may control the binding of TFs, where histones and TFs can compete for access to the DNA. These factors possess the ability to regulate the formation of the PIC (Lenhard et al., 2012) and may operate by recruiting other co-factors, such as Mediator, which can form a bridge between distal regulatory regions and the core promoter (Borggrefe et al., 2011). Sequence specific TFs may also recruit cofactors capable of remodelling nucleosome positioning or adding specific histone modifications associated with gene activation or repression (Engstrom et al., 2007; Forrest et al., 2014). Combinatorial occupancy of multiple TFs at those distal and proximal regulatory elements can lead to discrete and precise patterns of transcriptional activity (Lettice et al., 2012), suggesting that they can act cooperatively in regulating transcription (Whitfield et al., 2012).

1.4 Probing the three-dimensional organisation and the accessibility of the mammalian genome

Methods such as microscopy, in particular fluorescent *in situ* hybridisation (FISH), can clearly probe the organisation of chromosomes in the nuclear space, but have difficulties measuring multiple discrete contacts simultaneously and are limited in resolution and throughput. To overcome these barriers, a number of molecular methods have been developed. The seminal study by Dekker et al. (Dekker et al., 2002) first describing the Chromosome Conformation Capture (3C) method has sparked the development of a large number of 3C-derived genomics

methods. This technique and its derivates measure the averaged frequency at which two DNA fragments physically associate in the three dimensional (3D) space, based on their propensity to be cross-linked together. Consequently the initial step in 3C and 3C-derived methods is to freeze the DNA`s 3D structure by using formaldehyde as a fixing agent. Once interacting loci are cross-linked, restriction enzymes are used to fragment the chromatin. Most commonly, restriction enzymes with a six bp target sequence, such as HindIII or BglII, are used to cut the fixed chromatin. In addition, more frequent cutters, like AciI or DpnII, can be used (Comet et al., 2011; Miele et al., 2009). Subsequently the sticky ends of the obtained cross-linked DNA fragments are re-ligated under diluted conditions (Lieberman-Aiden et al., 2009; Schoenfelder et al., 2015a) or in the intact nucleus (Nagano et al., 2013; Rao et al., 2014) so that only covalently cross-linked fragments form ligation products. DNA fragments that are far away on the linear template, but co-localize in space, can be ligated to each other, thereby creating a chimeric one-dimensional cast of the 3D nuclear structure. These chimeric ligation products contain the information of not only where they originated from in the genomic sequence, but also where they reside, physically in the 3D structure of the genome. The way to establish the 3D conformation of a locus or chromosome is to measure the number of ligation events between non-neighboring sites. In 3C, this is done by semi-quantitative (Dekker et al., 2002) or quantitative (Miele et al., 2009) PCR amplification of selected ligation junctions, measuring the frequency of “one vs. one” interactions. Prior knowledge, or a strong hypothesis, of interacting genomic regions is required to choose the loci-specific primers required to amplify and quantify ligation junctions and is thus limited to the confirmation of suspected interactions rather than identifying novel interactions in an unbiased manner. The rapid progress of genome-scale methods such as microarrays and high-throughput next generation sequencing (NGS) has enabled the development of more unbiased 3C-related methods to analyze the three-dimensional organization of chromatin; these include 4C (“one-to-all” approach) (Simonis et al., 2006; Zhao et al., 2006), 5C (“many-to-many” approach) (Dostie et al., 2006) and Hi-C (“all-to-all” approach) (Lieberman-Aiden et al., 2009). In Capture-C, a standard 3C experiment is performed; however the 3C ligated fragments are sonicated and enriched for specific regions of interest with biotinylated oligonucleotides (Hughes et al., 2014), while Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) utilises a protein pull-down to achieve enrichment of spatial contact maps of the genome (Fullwood et al., 2009). A schematic overview of the different 3C derivatives is depicted in Figure 1.4.

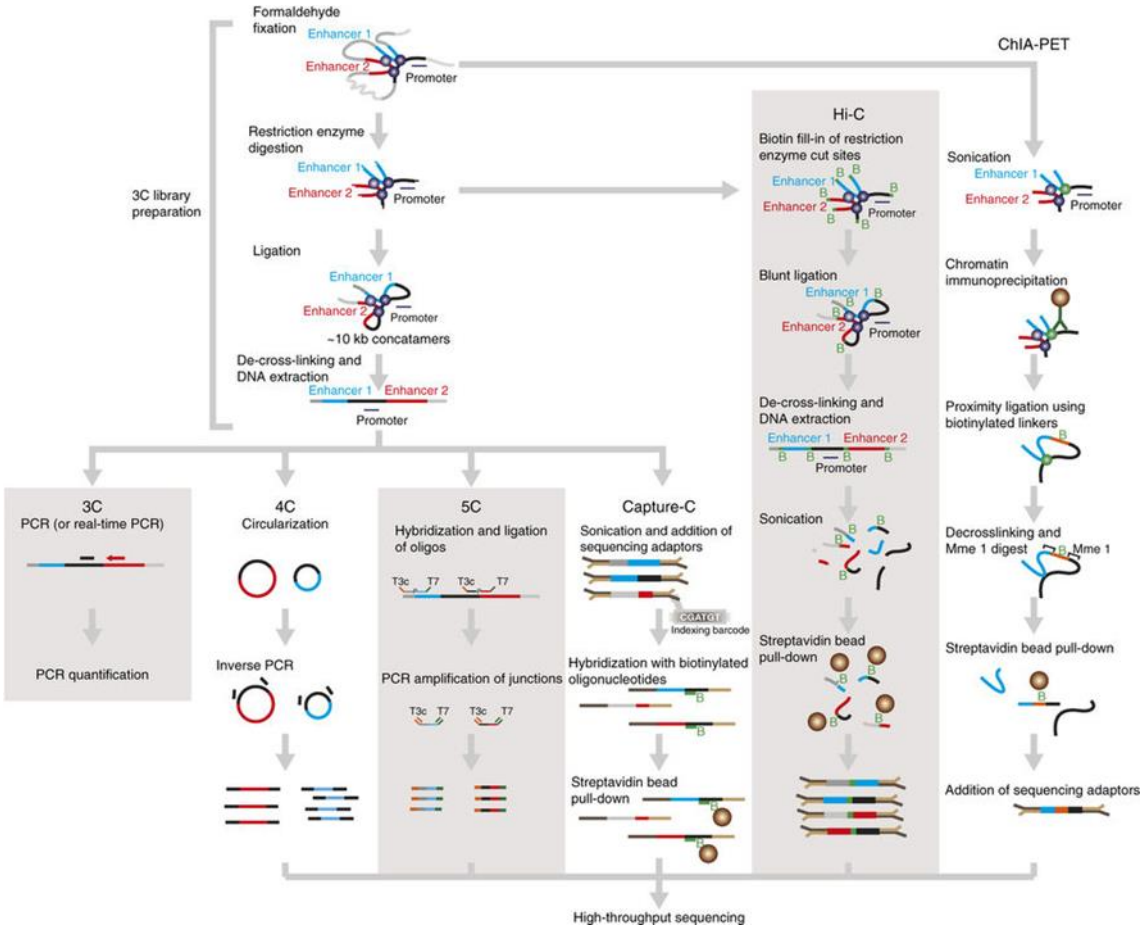


Figure 1.4 | Comparison of different 3C-based methodologies

Taken from (Davies et al., 2017) 3C libraries share the indicated steps and can then be interrogated by the specific steps in the 3C protocol shown subsequently.

Whilst 4C and 5C have been instrumental in advancing our knowledge and understanding of nuclear architecture and its role in genome regulation, (Andrey et al., 2013; de Wit et al., 2013; Pasquali et al., 2014; Simonis et al., 2006) Hi-C is unique in its ability to generate contact maps between all parts of the genome.

In Hi-C, the procedure for creating a 3C template is slightly adjusted. Before ligation, the restriction ends are filled in, using a nucleotide mix that contains a biotin-labeled nucleotide. Following a blunt end ligation, this biotinylated nucleotide now marks all ligations junctions in the Hi-C library. Subsequently, the DNA is purified and sheared, and a biotin pull-down is performed to ensure that only ligation junctions are selected for further analysis, such as paired-end deep sequencing after adapter ligation and PCR amplification. Reads are mapped back to the genome, and when a sequence read pair contains two different restriction fragments, this is scored as an interaction between those two fragments. From this, a matrix of ligation frequencies between all fragments in the genome can be constructed. The resolution of resultant Hi-C maps is determined by the restriction site density as well as depth of sequencing, and has increased from a scale of 1 Mb (Lieberman-Aiden et al., 2009) to single

kilobase resolution (Rao et al., 2014). For example, for a six-cutter like BglII, there are over 800,000 restriction sites in the mouse genome, resulting in a theoretical maximal genome-wide resolution of around 4 kb for a Hi-C experiment. However, this value is only theoretical, as a number of factors can influence this resolution. Local distribution of restriction sites varies between different genomic regions, resulting in different resolutions at different genomic locations. Further, a Hi-C library is composed of a massive variety of ligation products (up to 10^{11} unique pair-wise interactions between all 4 kb fragments; Belton et al., 2012), thus the number of sequence reads will ultimately determine the resolution of the interaction maps. Because of the quadratic nature of "all versus all" data, an increase in resolution by 10-fold requires a 100-fold increase in sequence depth and, therefore, the production of quality Hi-C data for large mammalian genomes requires the sequencing and mapping of several billion reads per sample. In order to reduce the number of sequencing reads required to generate contact maps of equivalent resolution for selected genomic regions, capture Hi-C was developed. An additional hybridisation selection of chosen loci, e.g. gene promoters (Schoenfelder et al., 2015a), specifically enriches Hi-C libraries for chosen loci and the DNA elements that they contact in 3D (Jager et al., 2015; Mifsud et al., 2015; Schoenfelder et al., 2015a), therefore increasing coverage and resolution.

Within the nucleoprotein structure of chromatin, epigenetic information is encoded and major insights have been gained from high-throughput, genome-wide methods by separately assaying the chromatin accessibility ("open chromatin") (Boyle et al., 2008), nucleosome positioning (Barski et al., 2007; Schones et al., 2008) and TF occupancy (Gerstein et al., 2012). ATAC-Seq captures open chromatin sites using a simple 2-step protocol and reveals the interplay between genomic locations of open chromatin, DNA binding proteins, individual nucleosomes and higher-order compaction at regulatory regions, thus providing a multi-dimensional portrait of gene regulation with nucleotide resolution (Buenrostro et al., 2013).
ATAC-Seq uses Tn5 transposase, loaded with sequencing adapters, to integrate these adapters into regions of accessible chromatin, whereas steric hindrance less accessible chromatin makes transposition less probable. Consequently, amplifiable DNA fragments suitable for high-throughput sequencing are preferentially generated at locations of open chromatin (Figure 1.5).

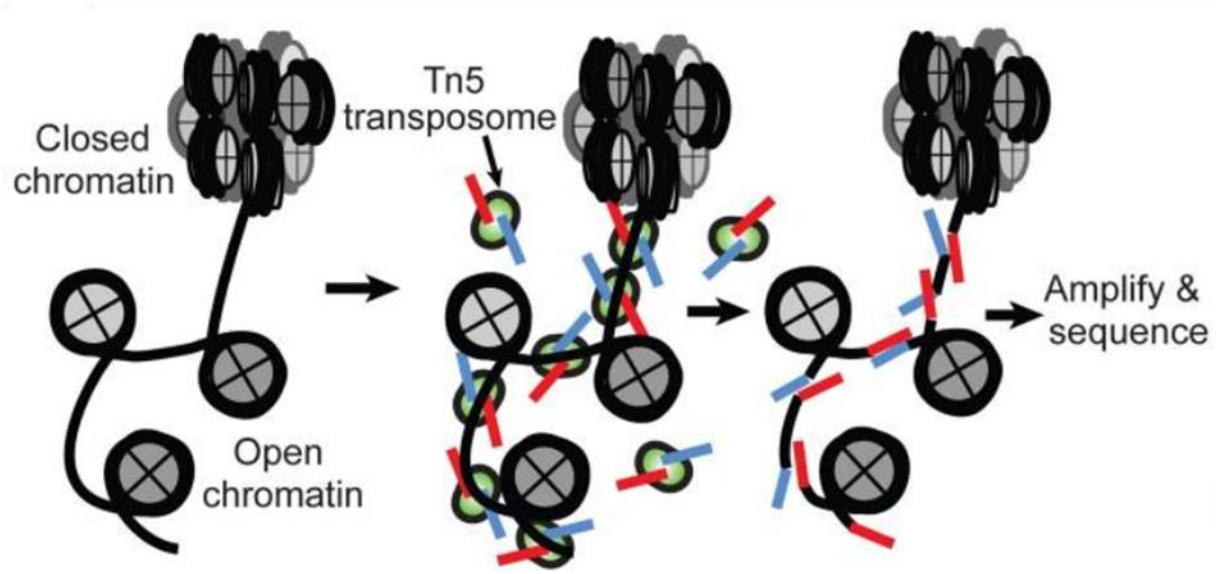


Figure 1.5 | ATAC-Seq reaction schematic

Transposase (green), loaded with sequencing adapters (red and blue), inserts only in regions of open chromatin (nucleosomes in grey) and generates sequencing library fragments that can be PCR amplified. Figure taken from (Buenrostro et al., 2013).

1.5 Three-dimensional organisation of the mammalian genome

Many processes and structures influence the 3D organisation of the genome. These include the relative positions of chromosomes in the nuclear space, long-range chromosomal associations and transcriptionally competent chromatin hubs. This section explores how the organisation of the genome in 3D contributes to gene expression.

1.5.1 Chromosome territories (CTs)

On the highest level of organisation, interphase chromosomes reside in discrete, almost mutually exclusive CTs (Cremer & Cremer, 2001), discovered by laser-UV micro-irradiation experiments (Cremer et al., 1982) and later confirmed by FISH experiments (Schardin et al., 1985). CTs can also be seen in more frequent intra-chromosomal (*cis*) chromatin interactions mapped by 3C based methods, compared to inter-chromosomal (*trans*) interactions (Dekker et al., 2013). In addition to occupying a defined territory within the nuclear space, small, gene-rich chromosomes tend to pair and co-localise to the nuclear interior (Bolzer et al., 2005; Cremer et al., 2001). Cell-type specific radial positioning of CTs within the nucleus has been reported (Parada et al., 2004), although the extent to which CT pairing and positioning are conserved through mitosis varies depending on the cell type analysed (Cremer et al., 2001). Individual genes are largely confined to their respective CT, however, in certain developmental contexts, such as the *Hox* gene activation (Chambeyron & Bickmore, 2004) and X-chromosome inactivation (Chaumeil et al., 2006), gene loci have been shown to loop out or move to the outer edges of their CTs and intermingle with stretches of DNA from different chromosomes.

The nuclear periphery, specifically the nuclear lamina, is linked to gene transcriptional silencing across the eukaryotic kingdom (Andrulis et al., 1998; Kosak et al., 2002; Pickersgill et al., 2006), and ectopic targeting of genetic loci to the nuclear envelope can induce transcriptional silencing in some cases, through the interaction of HP1 with B-type lamins, the major constituent of the nuclear envelope (Poleshko & Katz, 2014).

Spatial genome organization implies movement. The tissue-specific clustering of specific genomic elements requires that at some stage chromatin regions must move towards each other, in either a directed or a passive way. As cells exit mitosis and chromosomes de-condense, large-scale movements of chromatin domains have been observed (Thomson et al., 2004; Walter et al., 2003); these may result in the repositioning of chromosomal and sub-chromosomal regions to their generalized relative positions.

1.5.2 Hierarchical genomic domains

It has been described (Lieberman-Aiden et al., 2009) that chromosomes can be separated into two distinct chromatin compartments, namely the A and B compartments, based on their first or second principal component of Hi-C data. These two compartments have been found to correlate well with the chromatin state, including DNA accessibility, gene density, replication timing, GC content and histone marks. Thus, stretches of A-compartment DNA are interpreted as active, euchromatic regions, while B-type compartment DNA is interpreted as inactive, heterochromatic. The compartments show peculiar rates of decay of interactions with increasing genomic distance suggesting that the underlying compaction of chromatin belonging to each cluster is different. Loci separated by a certain linear distance belonging to the active compartment show a lower frequency of interactions than loci separated by the same linear distance belonging to the inactive compartment, pointing to the latter being more compacted (Lieberman-Aiden et al., 2009; Sexton et al., 2012).

The two genomic compartments have also been shown to display high plasticity, such that they change in different cell-types and biological conditions, matching large-scale changes in gene activity (Dixon et al., 2015). Individual compartment blocks tend to be in the order of 1-10 Mb in length, and, are thus easy to extract even in experiments with very low sampling. Furthermore, blocks of the active A-compartment tend to interact with other active stretches of DNA and inactive regions tend to cluster in 3D with other inactive B-compartments. It is important to note that while compartment signal is strong and easy to observe in large bins, the interaction frequencies at individual positions that have the same compartment type are low. Given that Hi-C measures a population average, it is likely that this pattern reflects a general, highly stochastic tendency of compartments to interact, rather than a set of deterministic interactions defined by individual loci.

Further, chromosomes appear to be folded as a hierarchy of nested chromosomal domains (Dixon et al., 2012; Sexton et al., 2012) and these are also thought to be involved in regulating genes, e.g. by limiting enhancer-promoter interaction to only those within a single chromosomal domain (de Laat et al., 2013; Gibcus & Dekker, 2013; Gorkin et al., 2014). Topological Associated Domain (TADs) represent linear evolutionary conserved sub-megabase self-interacting domains with shared epigenetic features (Dixon et al., 2012; Nora et al., 2012). Interactions within a given domain are more frequently observed than interactions between loci in different domains, even when those would be closer on the linear sequence. Thus, in the interaction matrix, TADs appear as square blocks of elevated interaction frequencies centred on the diagonal. Inter-domain interactions can occur, but preferentially between TADs of the same chromatin state, e.g. active with active and inactive with inactive. Additionally, changes in gene expression upon differentiation or external stimuli, such as IF treatment, are more likely to occur in the same direction for genes within a TAD than for genes in different TADs (Le Dily et al., 2014). The location of TAD boundaries are thought to be conserved between different cell types and even between different species, such as man and mouse, particularly within syntenic regions. Notably, the vertebrate insulator protein CTCF is enriched at a large subset of mammalian TAD boundaries (Dixon et al., 2012). Further, CP190, a critical contributor to the function of many *Drosophila* insulator proteins, is enriched at TAD boundaries in the fly (Sexton et al., 2012), suggesting an evolutionary conserved mechanism of TAD formation by insulator proteins. In particular, the Cohesin complex, which canonically, forms a ring around sister chromatids during mitosis (Peters & Nishiyama, 2012), has been shown to play a major role in organising DNA topology and affecting gene regulatory processes at the level of promoter-enhancer contacts. Additionally, Cohesin binding sites overlap significantly with CTCF binding sites genome-wide (Parelho et al., 2008; Wendt et al., 2008) many of which are conserved across cell types and species (Kim et al., 2007), leading to a model wherein CTCF-associated Cohesin localisation is thought to underlie the conservation of TAD boundaries across cell types and species, as hypothesised by Dixon et al (Dixon et al., 2012). It is also important to note that TAD boundaries are enriched in housekeeping genes and Short Interspersed Nuclear Elements (SINEs), but the role of transcription and SINEs in respect to TAD boundary formation remains elusive.

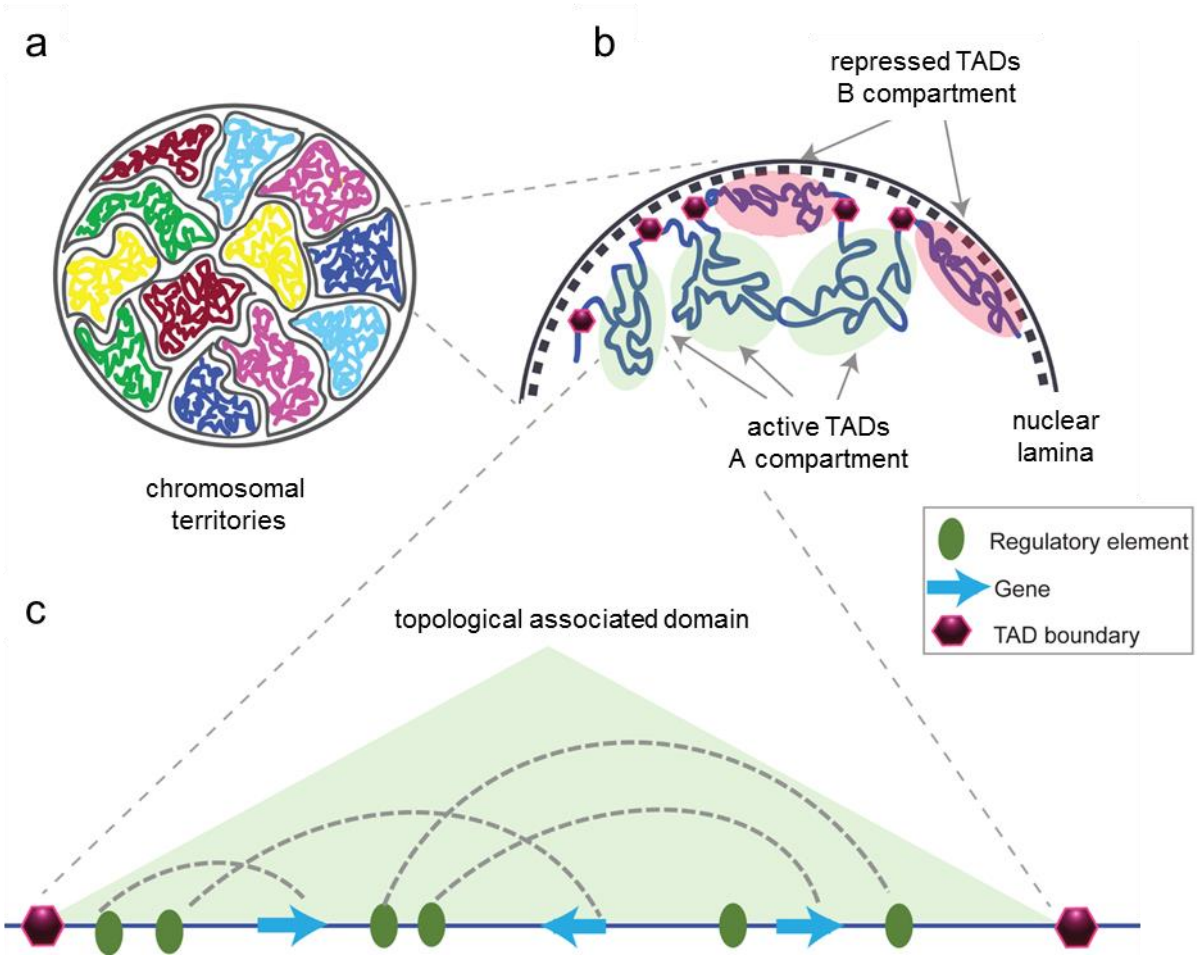


Figure 1.6 | Structural organisation of chromatin

(a) Interphase chromosomes of a diploid mammalian cell occupy distinct, almost mutually exclusive chromosomal territories. **(b)** Each chromosome is subdivided into A and B compartments, as found in Hi-C studies (Lieberman-Aiden et al., 2009). TADs with repressed transcriptional activity tend to be associated with the nuclear lamina (dashed inner nuclear membrane and its associated structures), while active TADs tend to reside more in the nuclear interior. TADs are flanked by insulating regions that are reducing interaction frequencies between different TADs, as determined by Hi-C, which are called TAD boundaries (purple hexagons). **(c)** An exemplary schematic of an active TAD with several interactions between distal regulatory elements (green oval) and genes (blue arrows) within the same TAD (green triangle). Figure modified from (Matharu & Ahituv, 2015)

1.5.3 Transcription factories

To fully understand the structure and function of the genome, it must be seen in the context of the nuclear proteome. Proteinaceous subcompartments are found within the nucleus and aid the control of chromatin dynamics and efficient functioning of related processes. One class of nuclear subcompartment which has come to light within the past thirty years is the transcription factory, seen as foci of hyper-phosphorylated RNA polymerase II scattered throughout the nucleus. The vast majority of genic transcription appears to take place at transcription factories (Osborne et al., 2004; Schoenfelder et al., 2010b) challenging the classical model of transcription found in many textbooks.

The term “transcription factories” was coined by Jackson et al. in 1993. Fluorescence microscopy was used to label the incorporation of bromouridine triphosphate (BrUTP) into nascent RNA; discrete foci of nascent transcription could then be seen within the nucleus which did not form in the presence of the RNA polymerase II inhibitor α-amanitin (Jackson et al., 1993). Further studies showed that these foci contained RNA polymerase II along with many other components required for transcription (Grande et al., 1997; Iborra et al., 1996).
The discovery of transcription factories has demanded a new model for the action of RNA polymerase II (Cook, 1999). The revised model proposes that instead of RNA polymerase II freely diffusing to active genes and tracking along the gene body, genes are recruited to transcription factories and are pulled through a stationary polymerase (Cook, 2002). It had been demonstrated that transcription is a discontinuous process with the frequency of nascent RNA transcription foci related to primary transcript RNA concentrations, suggesting that transcription occurs in bursts (Osborne et al., 2004). In a later paper, Osborne et al. showed that the immediate-early genes *Myc* and *Fos* are dynamically recruited to existing transcription factories within five minutes of B-cell stimulation, suggesting that the recruitment of genes to pre-existing transcription factories may be a method of transcriptional control (Osborne et al., 2007).

1.6 Rational and aims of the investigation

Our understanding of nuclear architecture and organisation has developed rapidly in the last twenty years, in concert with the development of techniques allowing ever larger and less biased studies. It is now widely accepted that nuclear structure can affect the function of the nucleus. These observations raise the possibility that the observed dramatic changes in nuclear architecture in certain diseases, such as viral infections and viral-induced carcinogenesis, may be advantageous for, and therefore actively promoted by, the infecting virus.
The first results chapter of this investigation provides a detailed, unbiased view on the transcriptional activity of mCMV infected cells. Studies in the past gained first useful insights, but were limited by their scope and by the techniques used. Here, I profile transcriptional changes in real-time in an unbiased manner, with an unprecedented temporal resolution. I hypothesised that by combining high temporal resolution expression profiling with accessibility data, a clear picture of the regulated pathways and the involved TFs will emerge.

Moreover, the dramatic phenotype of mCMV infection on the nuclear architecture, observed by microscopy, is rather striking. So far, a high-resolution genome-wide picture is missing and the scope of the global disruption of host cellular contacts is unknown. Furthermore, which regions of the host genome are compacted and how this impacts transcription remained elusive. I employed Hi-C and capture Hi-C methods to investigate the spatiotemporal changes

of the host and virus genomes during infection, and I integrated this data with the observed transcriptional changes.

The third aim of this investigation was to determine whether HPV16 integration in pre-malignant, unselected cells results in the aberrant expression of host genes. A number of studies have used advanced cervical cancer cell lines to demonstrate that HPV integration can cause a wide variety of somatic mutations, genomic amplifications and rearrangements resulting in the disruption of cellular genes. However, it remains unknown whether this phenomenon occurs as a result of all integration events or whether it is seen only in cells that are ultimately selected for. Furthermore, the role of short- and long-range interactions between the host and the integrated viral genomes in transcriptional dysregulation was unknown. To address these questions, I have used a panel of W12 cell integrant clones and assessed their chromosomal interaction profiles with the host genomic DNA.

Together, the experiments and analyses in this thesis provide an integrated view on the interplay between transcriptional response, chromatin accessibility and spatial chromatin architecture during viral infection, using mCMV and HPV16 as model viruses.

2 Methods

2.1 Cell culture and virus propagation

2.1.1 mCMV virus stock generation and titration

For the propagation of mCMV, NIH-3T3, cultured in DMEM supplemented with 5 % fetal calf serum and Pen/Strep, were used. Twenty cell culture dishes (diameter 14.5 cm) of 50 % confluent NIH-3T3 cells were infected at an MOI of about 0.1 with C3X R129 BAC-derived mCMV Smith strain. After 4-5 days, the supernatant of the infected cells was harvested and cell debris were removed by centrifugation at 4°C for 10 minutes at 10,000 g in a Sorvall centrifuge. The pellets were resuspended in 4 ml of medium and dounced 20 times. The obtained homogenate was then centrifuged for 10 min at 4°C and 10,000 g to remove remaining cell debris. The cell-free supernatants were then combined and virus particles were centrifuged for 3 h at 4°C and 28,000 g. The virus pellet was resuspended in 2 ml virus stock buffer (50 mM Tris-HCl, 120 mM KCl, and 5 mM EDTA) and dounced 5 times. The virus stock was aliquoted, flash frozen in liquid nitrogen and stored at 80°C.

To determine the viral titer, NIH-3T3 cells were plated in 48-well plates. Approximately 60 % confluent cells were infected in dilutions of virus stock ranging from 10^{-3} to 10^{-9} (diluted in DMEM). Centrifugal enhancement was used as described above to infect the cells. Cells were incubated for one hour at 37°C before the supernatants were removed and cells were overlaid with 500 µl carboxymethylcellulose medium, to restrict virus spread to cell-to-cell spread. After 4 days this medium was aspirated, cells were fixed in 10 % formaldehyde for 30 min at RT and subsequently stained with a 1 % crystal violet solution. Plaques were counted under a microscope and the virus titer was calculated. This was done using the following equation:

Virus titer (PFU / ml) = $\frac{Number\ of\ plaques}{d*V}$

Where, d = dilution

V = volume of diluted virus added to the plate

Carboxymethylcellulose-Medium (500 ml):

3.75 g carboxymethylcellulose (Sigma), 388 ml water, 25 ml FCS, 50 ml 10x MEM, 0.3 mg/ml L-glutamine, 2.5 ml Solution of non-essential amino acids (Gibco), 5 ml Penicillin/Streptomycin (10,00 U/ml), 24.7 ml 7.5 % $NaHCO_3$-solution

2.1.2 W12 keratinocyte cell line culture

All W12 derived clonal cell lines were grown in monolayer culture to mimic the basal layer of the epithelium. All cell lines were grown in complete culture medium: Glasgow Minimum Essential Medium (GMEM) supplemented with 10 % (w/v) Fetal bovine serum (FBS) (Sigma-

Aldrich, St. Louis, U.S.), 2 mM L-glutamine (Thermo Fisher Scientific, Loughborough, UK) and 100 U/mL of penicillin and 100 µg/mL of streptomycin (Thermo Fisher Scientific). Cells derived from the W12 keratinocyte cell line were co-cultured with X-ray irradiated G3T3 feeder cells (Todaro & Green, 1963) and grown in complete medium, as previously stated, with the addition of 10^{-10} M cholera toxin (Sigma-Aldrich), 0.5 µg/mL hydrocortisone (Sigma-Aldrich) and with 10 ng/mL epidermal growth factor (EGF) (Sigma-Aldrich) added 24 hours after seeding. All cells were tested fortnightly for mycoplasma contamination using Mycoplasma PlusTM PCR Primer Set (Agilent Technologies, Santa Clara, U.S.).

Cells were analysed at the lowest available passage (p) after cloning (typically p3 to p8), in order to minimise any effects of genomic instability caused by deregulated HPV16 oncogene expression. All clones analysed were episome-free. Formaldehyde crosslinking of 30 million cells was performed by supplementing standard EGF positive culture medium with formaldehyde to a final concentration of 2 % and was carried out for 10 min at room temperature. Crosslinking was quenched by the addition of ice-cold glycine to a final concentration of 125 mM. The adherent cells were scraped from the cell culture plates after crosslinking, collected by centrifugation (400 g for 10 minutes at 4˚C), and washed once with PBS (50 ml). After centrifugation (400 g for 10 minutes at 4˚C), the supernatant was removed, and the cell pellets were snap-frozen in liquid nitrogen and stored at -80˚C.

2.1.3 Murine fibroblast culture and virus infection

Murine NIH-3T3 fibroblasts (ATCC CRL1658) were cultured in DMEM (Gibco) supplemented with 5 % fetal calf serum and Penicillin-Streptomycin (100 U/mL final concentration, Invitrogen). Cells were split twice a week, tested fortnightly for mycoplasma contamination using Mycoplasma PlusTM PCR Primer Set (Agilent Technologies, Santa Clara, U.S.) and only passages 5 to 13 were used in experiments. Cells were seeded 20 h before experiments and grown overnight to 80 % confluence, followed by infection with C3X R129 BAC-derived mCMV Smith strain. Infection was performed at an MOI of 10 using centrifugal enhancement (30 min, 2000 rpm, room temperature), or at an MOI of 0.5. The time point just after centrifugation was considered as time point ‘0 min’ in all experiments. For harvesting, cells were scraped from the plates using cell scrapers and pelleted at 4°C for 5 min with 1600 rpm if needed, prior to further treatment/processing.

2.2 Imaging

2.2.1 Immunofluorescence imaging

NIH-3T3 were grown on coverslips to a density of ~50 % and were infected at an MOI = 10 as described in 2.1.2, using centrifugal enhancement. Cells were washed three times with 1x PBS prior to fixation. Cells were fixed and permeabilized using ice-cold methanol. Cells were

incubated for 2 h at RT with an anti-actin antibody coupled with Alexa Fluor (ab206277). DAPI stock solution was diluted to 300 nM using PBS and cells were incubated with the dilution for 5 min at RT. The sample was rinsed several time with PBS. The coverslips were mounted and imaged using the Zeiss 780 confocal microscope. Z-stacks of 1 µm were taken and visualized using the Imaris program.

2.2.2 Fluorescent *in situ* hybridization (FISH)

The method for performing 3D FISH with directly labelled DNA probes was implemented as described previously (Bolland et al., 2013) and performed by Emma Knight. For workflow see Figure 2.1.

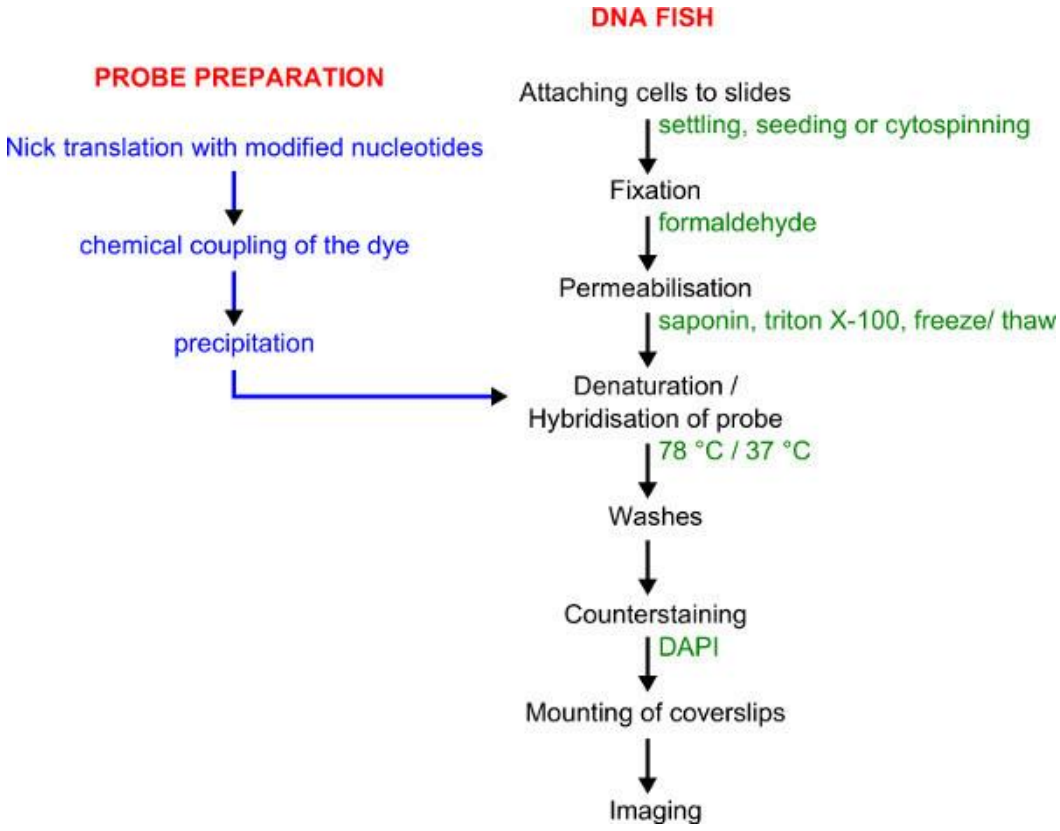


Figure 2.1| Workflow for probe labelling and DNA FISH

Figure taken from (Bolland et al., 2013).

W12 G2p11 cells (same passage number as used in SCRiBL) were resuscitated and grown in a 10 cm^2 tissue culture dish. Once at 70-80 % confluency the cells were trypsinised, washed in ice-cold PBS. The cells were then diluted to $5x10^5$/mL in 1x PBS. 20 µl of the cell suspension was then pipetted on to the centre of a polysine TM slide (VWR) and incubated at RT for 30 minutes to allow the cells to settle and adhere to the slide. Following incubation, the cells were fixed on the slide by submerging them into 4 % paraformaldehyde for 10 minutes. The fixation reaction was quenched with glycine and cells were then permeabilised in a 0.1 % saponin, 0.1 % Triton-X and 1x PBS solution at RT for 10 minutes. The slides were washed

once in 1x PBS and stored at -20 °C in 50 % glycerol in PBS until further use. High quality BAC DNA was extracted from the strains listed in Table 2.1 as described in 2.7.

Table 2.1 | Details of BAC clones used for 3D DNA FISH

BAC ID	Species, chromosome	Coordinates	Supplier	Alexa Fluor
RP11-467N14	Human Chr 5	51,676,020-51,873,551	Thermo Fisher	647
CTD-2015C9	Human Chr 5	53,473,886-53,584,235	Thermo Fisher	555
pSP64-HPV16	HPV16	0-7904	Cinzia Scarpini	488

Nick translation, coupling of the fluorescence dye and probe precipitation were performed according to the published protocol (Bolland et al., 2013). DNAse I was used to nick DNA in the presence of Polymerase I, which elongates the 3`ends of the nicks and replaces existing nucleotides with new ones, thereby “translating” the nick and thus providing the opportunity to incorporate labelled nucleotides (aminoallyl–dUTP). Fluorescent labelling of the probe is achieved by chemical coupling of the amine-reactive dye with the aminylated probe (see Table 2.1).

The cells were permeabilised using repeated cycles of flash freezing and thawing using liquid nitrogen. Probes and cells were combined, denatured at 78°C for 2 min and hybridised at 37°C overnight. Slides were washed and stained with DAPI solution for 2 min at RT. Slides were fixed again with 3.7 % formaldehyde for 5 min, quenched with glycine and washed with PBS, before mounting and sealing of the cover slips.

All FISH slides were analysed at the Babraham Institute imaging facility, by Olga Mielczarek (PhD student, Corcoran Lab, Babraham Institute). Successful probe hybridisation was determined by first looking at the slides using an Olympus FV1000 confocal microscope. Once a successful FISH reaction had been confirmed, the slides were transferred to a MetaSystems Metacyte connected to Zeiss Axio Imager Z2 microscope for analysis. For this experiment, a 3-probe assay of wavelengths 488 nm (green), 555 nm (red) and 647 nm (far red) was used. The Metafer 4 v3.11.2 software was used to perform an automated analysis of fluorescence signals including the identification of the number of fluorescent spots per cell and distance between signals. Roughly, 1500 cells per slide were captured using the Metacyte; cells containing human BAC probes x 2 and HPV probe x 1 were discarded from the subsequent analysis of the 3D distances between probes. The x, y and z coordinates between all three probes were exported and analysed using a customised Perl script (Felix Krueger, Babraham Institute).

2.3 ATAC-Seq library preparation

Infected and non-infected NIH-3T3 were harvested and pelleted as described (Section 2.1.3). ATAC-Seq was performed mainly as described before (Buenrostro et al., 2013), starting with ~100,000 cells. Cells were washed in PBS, spun down at 4°C for 5 min at 272 g and lysed on ice for 15 min in 50 µl ice-cold lysis buffer (10 mM TrisHCl pH7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 % Igepal). Nuclei were pelleted at 600 g for 10 min at 4°C and washed once in ice-cold lysis buffer. After spinning down the washed nuclei, as described in the step before, nuclei were resuspended in 22.5 µL nuclease-free water, 25 µL 2x TD buffer and 2.5 µL transposase (TDE1) from the Nextera DNA Sample Preparation Kit (Illumina, FC-121-1030) and incubated at 37°C for 30 min. Following DNA purification on a MinElute column (Qiagen), libraries were amplified using the Nextera DNA Sample Preparation Kit and the index primers from the Nextera Index Kit (Illumina, FC-121-110). See Table 2.2 for the primer combinations of the individual time points. Libraries were amplified using 10 PCR cycles, based on the published protocol (Buenrostro et al., 2013) with the following conditions: 72°C for 5 min; 98°C for 30 s; thermocycling at 98°C for 10 s, 63°C for 30 s and 72°C for 3 min; and a final extension at 72°C for 5 min. After removing residual primers with the PCR clean-up kit (Qiagen), the library concentrations and size distributions were assayed with the KAPA Library Quantification Kit (KAPA Biosystems) and the Agilent Bioanalyzer 2100. The KAPA kit polymerase is less efficient above 1 kb; however, this should also reflect the efficiency at which different length fragments are amplified during the sequencing process. Bioanalyzer profiles were checked for an insert size distribution with ~200 bp periodicity. Libraries were sequenced by Cambridge Genomic Services (CGS) with 75 bp paired-end on a NextSeq 500 platform (Illumina) using the high output kit with two indexing reads on one lane per replicate.

Table 2.2 | Illumina Nextera barcodes used for ATAC-Seq

ATAC-Seq libraries were amplified using the Nextera Index kit with the below indicated primer combinations, resulting in dual indexing.

P7 primer	5`-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3`
P5 primer	5`-AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGCAGCGTC-3`

Time point	Replicate	i7 name	i7 sequence	i5 name	i5 sequence
Mock	1	702	CTAGTACG	501	TAGATCGC
1 hpi	1	702	CTAGTACG	502	CTCTCTAT
2 hpi	1	704	GCTCAGGA	503	TATCCTCT
4 hpi	1	704	GCTCAGGA	504	AGAGTAGA
6 hpi	1	706	CATGCCTA	501	TAGATCGC
12 hpi	1	702	CTAGTACG	502	CTCTCTAT

19 hpi	1	703	TTCTGCCT	503	TATCCTCT
24 hpi	1	704	GCTCAGGA	504	AGAGTAGA
48 hpi	1	705	AGGAGTCC	504	AGAGTAGA
Mock	2	701	TCGCCTTA	504	AGAGTAGA
1 hpi	2	702	CTAGTACG	504	AGAGTAGA
2 hpi	2	703	TTCTGCCT	504	AGAGTAGA
4 hpi	2	704	GCTCAGGA	504	AGAGTAGA
6 hpi	2	705	AGGAGTCC	504	AGAGTAGA
12 hpi	2	706	CATGCCTA	504	AGAGTAGA
19 hpi	2	704	GCTCAGGA	503	TATCCTCT
24 hpi	2	705	AGGAGTCC	503	TATCCTCT
48 hpi	2	706	CATGCCTA	503	TATCCTCT
Mock	3	701	TCGCCTTA	501	TAGATCGC
1 hpi	3	701	TCGCCTTA	502	CTCTCTAT
2 hpi	3	701	TCGCCTTA	503	TATCCTCT
4 hpi	3	703	TTCTGCCT	503	TATCCTCT
6 hpi	3	704	GCTCAGGA	501	TAGATCGC
12 hpi	3	704	GCTCAGGA	502	CTCTCTAT
19 hpi	3	705	AGGAGTCC	501	TAGATCGC
24 hpi	3	705	AGGAGTCC	502	CTCTCTAT
48 hpi	3	706	CATGCCTA	502	CTCTCTAT

2.4 4-thiouridine-tagging

2.4.1 Metabolic labelling of newly transcribed RNA

Experiments were performed in 6-well dishes to allow for the infection of cells with mCMV using centrifugal enhancement. RNA labeling was started by adding 200 µM 4-thiouridine (4sU, Carbosynth, # NT06186) to cell culture media for 1 h at different time points of infection. At the end of labeling, the medium was aspirated and total cellular RNA was isolated using 1 ml Trizol reagent (Invitrogen) per 6-well, resulting in 6 ml of Trizol sample per 6-well plate. This can be frozen and stored at -20°C for several weeks. After defrosting, 1.2 ml Chloroform (0.2 ml per 1 ml Trizol) was added and the solution and vigorously shaken for 15 s followed by incubation at room temperature for 3 min. The phases were separated by a centrifuging step at 4°C, 13,000 g for 15 min and the upper aqueous phase (containing the RNA) was transferred to a new 15 ml falcon tube. Precipitation was done by adding ½ of the reaction volume of RNA precipitation buffer (1.2 M NaCl, 0.8 M NaCitrtate) and ½ of the volume of isopropanol, i.e. 1.5 ml of each. After mixing well and incubation at room temperature for 10 min, the RNA was pelleted at 4°C and 13,000 g for 10 min. After the supernatant was removed the pellet was

washed with 75 % ethanol. After spinning at 4°C and 13,000 g for 10 min the ethanol was removed at the pellet was allowed to air dry. RNA was dissolved in 100 µl TE, heated to 65°C for 10 min and the concentration was measured using Nanodrop.

2.4.2 Biotinylation of newly transcribed RNA

Biotinylation of newly transcribed RNA was performed starting with 100 µg of total RNA in a total volume of 1 ml. 200 µl of MTSEA biotin-XX (Biotium #90066; 12.5 µg/mL), 100 µl 10x Biotinylation buffer (100 mM HEPES pH 7.4, 10 mM EDTA) and the corresponding amount of RNA were added, made up to 1 ml, and incubated with rotation under protection from light for 1.5 h at room temperature. Unincorporated biotin was removed in two chloroform extraction steps. An equal volume (1 ml) of Chloroform was added, mixed vigorously and incubated for 3 min. Phase separation was performed by spinning at 4°C and 20,000 g for 5 min. To reduce RNA loss the second extraction was done using Phase Lock Gel Heavy tubes (2 ml, Eppendorf). RNA was precipitated by adding 100 µl 5 M NaCl and an equal volume of isopropanol (1 ml) and pelleted at 4°C and 20,000 g for 20 min. After washing with 1 ml of 75 % ethanol and further spinning at 4°C 20,000 g for 10 min, the pellet was dried and resuspended in 100 µl TE.

2.4.3 Separation of labeled and unlabeled RNA

After heating the RNA to 65°C for 10 min, the labeled RNA was separated from the unlabeled RNA using Streptavidin-coated magnetic beads. For this purpose, 100 µl RNA were incubated with 100 µl µMac streptavidin beads with rotation for 15 min at room temperature. During this, µMacs columns were placed into magnetic stands and equilibrated with 0.9 ml RNA washing buffer (100 mM Tris pH 7.5, 10 mM EDTA, 1 M NaCl, 0.1 % Tween20). After applying beads to the columns, they were washed three times with 0.9 ml 65°C washing buffer, followed by three times with 0.9 ml room temperature washing buffer. Elution took place with two times 100 µl 100 mM DTT. By adding 200 µl of isopropanol and 1 mg glycogen, the newly transcribed RNA was precipitated. After pelleting it at 4°C and 20,000 g for 20 min, the RNA was resuspended in 40 µl TE and the concentration was measured using Nanodrop.

2.4.4 Strand specific RNA-Seq library preparation

Opposing strand specific libraries were generated by BGI Hong Kong without rRNA-depletion. First, RNA was fragmented by adding First Strand Master Mix (Invitrogen). First-strand cDNA was generated by First Strand Master Mix and Super Script II reverse transcription (Invitrogen; 25°C for 10 min, 42°C for 50 min, 70°C for 15 min). The product was purified with Agencourt RNAClean XP Beads, then Second Strand Maser Mix, dATP, dGTP, dCTP and dUTP were added and second strand cDNA was synthetised for 1 h at 16°C. Purified cDNA was end repaired for 30 min at 30°C, A-tailed and TruSeq Illumina compatible sequencing adapters

were ligated. Samples were digested with Uracil-DNA glycosylase prior to PCR amplification. Library concentrations were assessed by the Agilent Bioanalyzer 2100 and by qPCR. Libraries were sequenced by BGI Hong Kong with 100 bp paired-end sequencing on a HiSeq X Ten platform (Illumina), using one lane per replicate.

2.5 Hi-C library generation

The method described in the following section is based on the Nagano/Schoenfelder Hi-C protocol (Schoenfelder et al., 2015a), a modified version of the original method (Belton et al., 2012).

W12 derived cell lines were cross-linked as described in 2.1.2 NIH- 3T3 cells were cultured, infected and harvested as described in 2.1.3 resulting in 1.44 x 10^7 cells per time point. Cells were initially resuspended in 37 ml DMEM supplemented with 10 % FBS. To fix the cells, 43 ml of 16 % formaldehyde stock solution (Agar Scientific) was added to a final concentration of 2 % and incubated with rotation at room temperature for exactly 10 min. After formaldehyde cross-linking, the remaining formaldehyde was quenched by adding 6 ml 1 M ice cold glycine and incubated for 5 min at room temperature, followed by incubation on ice for 15 min. By centrifugation at 4°C and 400 g for 8 min, the cells were pelleted and then washed with 50 ml ice cold PBS. After removing the PBS as described in the previous step, the cell pellets were frozen in liquid Nitrogen and stored at -80°C.

Cells were thawed on ice, and then lysed on ice for 30 minutes in 50 ml freshly prepared ice-cold lysis buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 0.2 % Igepal CA-630, one protease inhibitor cocktail tablet (Roche complete, EDTA-free)). Following the lysis, nuclei were pelleted (650 g for 5 minutes at 4˚C), washed once with 1.25x NEBuffer 2 or 3, and then re-suspended in 1.25x NEBuffer 2 (for MboI digestion) or 1.25x NEBuffer 3 (for BglII digestion) to make aliquots of 5-6 million cells for digestion. SDS was added (0.3 % final concentration) and the nuclei were incubated at 37˚C for one hour (950 rpm). Triton X-100 was added to a final concentration of 1.7 % and the nuclei were incubated at 37˚C for one hour (950 rpm). Restriction digest was performed overnight at 37˚C (950 rpm) using 800 units MboI (NEB) or 1000 units BglII per 5 million cells. Restriction fragment ends were filled in using Klenow (NEB) with dCTP, dGTP, dTTP and biotin-14-dATP.The blunt-ended DNA was ligated following the in-nucleus ligation protocol as described previously (Nagano et al., 2015), or following the in-solution ligation protocol (Schoenfelder et al., 2015a), both with minor modifications. For in-solution ligation, 86 µl of 10 % SDS were added to each of the tubes containing DNA. Tubes were incubated at 37°C for 1 h with rotation and then immediately placed on ice. During this incubation 15 ml falcons, one for each 5 million cells starting material tube, were prepared, each with 745 µl 10 % Triton X-100, 820 µl 10x ligation buffer (NEB), 80 µl 10 mg/ml BSA and 5.965 ml water. For Hi-C blunt-end ligation, 50 µl 1 U/µl T4 DNA ligase (Invitrogen) were added

to each tube and incubated at 16˚C for 4 hours, followed by 30 min at room temperature. For the in-nucleus ligation, prior to ligation, excess salts and enzymes were removed by centrifugation (600 g for 5 minutes at 4˚C) and the cell pellet was re-suspended in 995 µl of 1x ligation buffer (NEB) supplemented with BSA (100 µg/mL final concentration). The ligation was carried out using 2000 units of T4 DNA ligase (NEB) per 5 Mio starting material of cells, at 16˚C for 4 hours, followed by 30 min at room temperature. After the ligation, chromatin was de-crosslinked overnight at 65˚C in the presence of proteinase K (Roche), purified by phenol and phenol-chloroform extractions, precipitated with ethanol and sodium acetate and re-suspended in TLE (10 mM Tris-HCl pH 8.0; 0.1 mM EDTA). The DNA concentration was measured using the Quant-iT PicoGreen assay (Life Technologies).

Hi-C ligation efficiency controls were conducted to assure that libraries meet quality metrics, before proceeding. Therefore, several PCRs were performed to amplify known short- and long-range interactions. Furthermore a PCR digest assay was done. For all PCRs the HotStar Taq DNA Polymerase Kit (Qiagen) was used according to the manufacturer's instruction, using 250 ng of template in a total volume of 25 µl per reaction and 36 PCR cycles. For restriction digest assay, 700 ng of a Hi-C short-range product were either mock digested, digested with one restriction enzyme using 20 units or digested using two restriction enzyme with 20 units each. PCR and digestion products were separated by agarose gelelectrophoresis and were visualized.

To remove of biotin moieties from non-ligated fragment ends, 40 µg of Hi-C library DNA were incubated with T4 DNA polymerase (NEB) for 4 hours at 20˚C, followed by phenol/chloroform purification and DNA precipitation overnight at -20˚C. DNA was sheared to an average size of 400 bp using the Covaris E220 (settings: duty factor: 10 %; peak incident power: 140 W; cycles per burst: 200; time: 55 seconds). End-repair of the sheared DNA (using T4 DNA polymerase, T4 DNA polynucleotide kinase, Klenow (all NEB) was followed by dATP addition (Klenow exo-, NEB) and a double-sided size selection using AMPure XP beads (Beckman Coulter) to isolate DNA ranging from 250 to 550 bp. Biotin-marked ligation junctions were immobilised using MyOne Streptavidin C1 Dynabeads (Invitrogen) in binding buffer (5 mM Tris-HCl pH 8.0, 0.5 mM EDTA, 1 M NaCl) and after stringent washing in the same buffer at 55°C for 10 min ligated to the custom SCRiBL adapter using 1600 units of T4 DNA ligase (NEB) for 2 hours at room temperature. These adapters were generated by annealing SCRiBL_adapter_1 and SCRiBL_adapter_2 (Table 2.5). The immobilised Hi-C libraries, subsequently subjected to enrichment, were amplified using the custom primers PE_PCR_1.0.33 and PE_PCR_2.0.33 and the Phusion polymerase (New England Biolabs) with 7-9 cycles. Final Hi-C libraries ready for sequencing were amplified using the TruSeq universal primer and a barcode specific reverse compliment (see Table 2.5 for sequences and Table 2.4 for the barcodes used per

library). After PCR amplification, the Hi-C libraries were purified with AMPure XP beads (Beckman Coulter). Quantity and integrity of the Hi-C libraries was determined by Bioanalyzer profiles (Agilent Technologies).

2.6 Genomic DNA library generation

Cells were thawed, lysed and nuclei were isolated as described above (2.5). Nuclei from 5-6 million cells were treated with SDS and Triton X-100 as described for the generation of Hi-C libraries. All Hi-C specific steps, such as MboI digestion, restriction fragment end fill-in, blunt end ligation and the removal of biotin from un-ligated restriction fragment ends were mock performed by replacing the respective enzymes with an equal amount of water. All other steps were performed as described for the generation of the Hi-C libraries. The biotin-streptavidin pull down was omitted and, therefore, the ligation of the custom sequence adapters was done in solution by adding 4 µl adaptors (30 µM) and 1600 units T4 DNA ligase (NEB). The ligation was carried at for 2 hours at room temperature on a rotating wheel in 1x ligation buffer (NEB).

Test PCRs on 1/20th of the library were run using the custom primers PE_PCR_1.0.33 and PE_PCR_2.0.33 and Phusion polymerase (New England Biolabs), in order to titrate the number of cycles needed for the final amplification.

Pre-capture PCR amplification was carried out using the custom primers PE_PCR_1.0.33 and PE_PCR_2.0.33 with 7-8 cycles utilising the Phusion polymerase (New England Biolabs). The amplified libraries were purified with AMPure XP beads (Beckman Coulter) and the quantity and the quality was assessed by Bioanalyzer profiles (Agilent Technologies). See Figure 2.2 for primer strategy outline.

2.7 BAC DNA preparation

BACs covering the mouse genome were ordered from Invitrogen as Glycerol stocks and stored at -80°C. All BACs used for FISH can be seen in Table 2.1 and BACs used for SCRiBL are listed in Table 2.3. To culture the bacteria, a sterile inoculation loop was dipped into the glycerol stock and then transferred to 5 ml LB-medium containing 15 µg/ml chloramphenicol. This was incubated for 3 h at 32°C, then added to 100 ml LB-medium containing the same concentration of chloramphenicol and incubated at 32°C overnight. The bacteria were pelleted at 6,000 g at 4°C for 15 min using a Sorvall centrifuge. BAC DNA isolation was done using the NucleoBond BAC 100 Kit (Macherey Nagel) according to the manufacturer`s protocol. The obtained DNA pellet was finally resuspended in 100 µl H_2O and the concentration was measured using Nanodrop. BAC DNA was stored at 4°C for short-terms and at -20°C for long-term storage.

LB-medium (1 L):

10 g Bacto Trypton, 5 g Bacto yeast extract, 5 g NaCl

2.8 SCRiBL bait generation for large region capture

An equimolar mix of the 24 mouse genomic BACs (Table 2.3), resulting in ~10 µg DNA, was digested with BglII. The mCMV BAC DNA was kept and treated separately. Thereof, 5 µg were used for BglII digest. Each of these DNA mixtures was digested over night at 37°C with 200 U BglII per 5 µg DNA in a total volume of 200 µl 1x NEB buffer 3 each. The digested DNA was extracted with phenol:chloroform followed by only chloroform using maxtract tubes. After adding 0.1 V 3 M Sodium acetate (NaOAc) and 2.5 V 100 % ethanol, the DNA was precipitated for 2 h at -20°C followed by 30 min at -80°C. The DNA was pelleted at 14,000 rpm and 4°C for 20 min, washed with 500 µl 70 % ethanol and resuspended in 10 mM Tris pH 7.5 (1 µl/µg of initially used µg of DNA). The concentration was measured with Nanodrop and the digest efficiency was estimated on a 0.8 % agarose gel.

To later enable *in vitro* transcription, T7 promoter adapter were ligated. The adaptors were prepared by mixing 100 µl of both Lou T7 BglII and Lou T7 BglII rev primers (100 mM) with 300 µl oligo annealing buffer (10 mM Tris pH 8, 50 mM NaCl, 1 mM EDTA), heating them to 95°C and cooling them down to 4°C with ΔT/min=1°C. To ligate the T7 promotor adaptors to the BglII overhangs of the digested BAC DNA, two reactions of the BAC mixture each with 10 µg and one reaction of the digested mCMV BAC DNA containing 1.4 µg DNA were set up and incubated with a 15-fold molar excess of adaptor (3.1 µl / 10 µg DNA), 2 µl (400 U/µl) T4 DNA ligase (NEB) in a total volume of 50 µl for 16 h at 16°C. The T4 DNA ligase was inactivated at 65°C for 10 min.

For sonication the 50 µl reactions were made up to 134 µl with 84 µl 1x T4 DNA ligase buffer (NEB). Four µl per sample were kept as pre-sonication sample for a gel and the remaining 130 µl were transferred to a Covaris microTUBE. The following conditions were used to generate fragments with a peak at 200 bp: duty factor 10 %, Peak Incident Power (w) 175 and 200 Cycles per Burst for 250 sec.

The sheared ends were repaired by adding 3.75 µl 10mM dNTPs, 3 µl 10x NEB ligase buffer, 12 µl H_2O, 5 µl (15U) T4 DNA Polymerase (NEB), 5 µl (50U) Polynucleotide Kinase (NEB) and 1 µl (5U) Klenow (NEB) followed by incubation at 25°C for 30 min. This was followed by a purification using the Qiagen PCR purification kit. Final elution was done with 50 µl H_2O. The DNA was run on a 1.5 % agarose gel and fragments from 180-300 bp were excised from the gel, purified using the Zymoclean Gel DNA recovery kit (Zymogen) and eluted in 30 µl H_2O. To get rid of the remaining agarose contaminations, SPRI bead purification was performed and the DNA was finally eluted in 21 µl H_2O.

In vitro transcription was performed using the T7 Maxiscript kit (Ambion) according to the manufacturer`s instruction with 800 ng DNA of the BAC mixture in a total volume of 100 µl. For

the mCMV DNA the reaction was down-scaled to a total volume of 40 µl using 120 ng of DNA. The reactions were incubated for 18 h at 37°C. The samples were treated with 1 µl Turbo DNase per 20 µl of reaction and incubated at 37°C for 15 min. The DNase was inhibited by adding the same amount of 0.5 M EDTA. The samples were made to 75 µl either by adding H_2O or splitting them into two and then were purified with G50 columns (Roche). The volume was made up to 200 µl, the RNA was extracted with acid (pH 4) phenol:chlororform and precipitated by adding 80 µl Ammonium acetate (NH_4OAc) and 700 µl 100 % ethanol and incubation over night at -20°C. RNA was pelleted, washed and finally resuspended in 11 µl H_2O.

Table 2.3 | BACs used for murine SCRiBL bait generation

BACs were ordered from Invitrogen and checked for the integrity by either restriction enzyme digest and/or PCR. Given coordinates are mm9 and were lifted over using the NiceLiftOver tool.

gene	BAC	bp	Chr.	start	end	class
Trp53inp2	RP23-87I24	250,753	2	155,184,799	155,435,552	up
CD34	RP23-82P18	250,094	1	196,724,427	196,974,521	down
Csf1	RP23-15F10	236,593	3	107,393,103	107,629,696	down
Egr3	RP23-359F5	230,540	14	70,401,784	70,632,324	down
Hook2	RP23-359C11	228,786	8	87,354,934	87,583,720	up
Hss	RP23-141E23	227,312	13	23,527,157	23,754,469	control
Hba	RP23-291N7	226,829	11	32,002,389	32,229,218	control
Azi1	RP23-50O10	220,478	11	119,846,518	120,066,996	up
Shpk	RP23-390G23	217,417	11	72,903,617	73,121,034	up
Nfkbia	RP23-119C24	216,250	12	56,493,833	56,710,083	NF-κB
Nanog	RP24-464B14	213,743	6	122,553,129	122,766,872	control
myc	RP23-442F1	207,374	15	61,741,905	61,949,279	down
Nfkbie	RP23-346N18	207,123	17	45,584,992	45,792,115	NF-κB
Aff3	RP23-424F5	206,928	1	38,564,008	38,770,936	up
Ifit1	RP23-111N1	205,719	19	34,563,196	34,768,915	IF
Ppfia3	RP23-193A10	203,877	7	52,532,521	52,736,398	up
Ifitm3	RP23-354L18	202,949	7	148,116,172	148,319,121	IF
cluster	RP24-66A10	198,690	13	22,659,336	22,858,026	control
Cxcl10	RP23-447A7	195,272	5	92,680,975	92,876,247	IF
Hbb	RP24-344M21	190,217	7	110,840,403	111,030,620	control
Fn1	RP23-236P18	178,653	1	71,574,977	71,753,630	down
Dusp5	RP23-146L13	176,027	19	53,568,582	53,744,609	down
Megf10	RP23-472D7	168,572	18	57,257,451	57,426,023	up

Tnfaip3	RP23-361I11	161,272	10	18,678,005	18,839,277	NF-κB
mCMV	C3X R129	238,500	mCMV	0	238,500	NA

2.9 SCRiBL bait generation for HPV16 capture from Hi-C libraries

120-mer capture RNA baits were complementary designed to both ends of MboI restriction fragments overlapping the HPV-16 genome. Requirements for target sequences were as follows: GC content between 25 % and 65 %, no more than two consecutive Ns within the target sequences, and the maximum distance to an MboI restriction site was 330 bp. For short MboI fragments, where 120-mer RNA baits originating from both ends would have overlapped (potentially interfering with optimal hybridization to Hi-C libraries), only the coding strand was used for capture RNA bait design, and if necessary the baits were trimmed to a minimum length not shorter than 97 nt. This resulted in the design of 16 RNA bait sequences covering the MboI restriction fragment ends of the entire HPV-16 genome, with the exception of two fragments too short (18 and 63 bp, respectively) for capture RNA bait design.

DNA sequences encoding for the 16 RNA bait sequences, with different restriction enzyme sites at each fragment end (5' BglII and 3' HindIII or SpeI, Figure 5.4) were ordered as two gBlocks (Integrated DNA Technologies, sequence in appendix) and cloned into plasmid vectors using the Zero Blunt® TOPO® cloning kit with One Shot® TOP10 chemically competent cells according to the manufacturer's instructions. Following overnight incubation, plasmid DNA was isolated using the QIAprep spin Miniprep kit (Qiagen) according to the manufacturer's instructions. DNA was eluted in 50 µl Elution Buffer. To isolate the gBlock fragments from the cloning vector, digestion reactions with the restriction enzyme EcoRI were performed. Therefore, 8.5 µg of each gBlock 1 and 2 were incubated with 12 µl 10x CutSmart Buffer and 6 µl EcoRI HiFi enzyme (both New England Biolabs) for 2 hours at 37 °C. After incubation the mixtures were run on a 1 % agarose gel and the bands containing the desired insert (~1 kb) were cut out and purified using the QIAquick Gel extract kit (Qiagen) according to the manufacturer's instructions. The DNA was eluted in 50 µl H_2O per column and quantified by Nanodrop.

DNA from both gBlocks was then further digested to release the fragments specific to the HPV16 genome MboI fragment ends. The DNA from gBlock1 was incubated with restriction enzymes BglII (10 U/µl) and HindIII (20 U/µl), whereas DNA from gBlock2 was incubated with BglII (10 U/µl) and SpeI (10 U/µl). Each fragment contained a single BglII cut site and enabled side specific ligation of a T7 promoter sequence adapter. T7 adapters were generated by annealing T7_promoter_adapter_1 and T7_promoter_adapter_2 (Table 2.5); 20 µl each of both forward and reverse primers (100 µM) were mixed with 60 µl oligo annealing buffer (10 mM Tris pH 8.0, 50 mM NaCl, 1 mM EDTA) and placed in a PCR machine at 95 °C for 5 minutes. Following this initial incubation, the temperature was decreased at a rate of 1°C per

minute to 4°C. Digestion with both restriction enzymes and adapter ligation was carried out in one reaction simultaneously in the presence of BamHI (20 U/µl) to each reaction in order to cut any unspecific adapter-adapter products that may be present. Two reactions containing 700 ng gBlock1 or 850 ng gBlock2 DNA, 30 units BglII each, 100 units BamHI each, 5-fold molar excess of pre-annealed T7 promoter adapters, and either 80 units HindII (NEB) or 40 units SpeI (NEB) were incubated at 37 °C for 2 hours in 1x T4 DNA ligase buffer (NEB). Following this incubation 1200 units T4 DNA ligase (NEB) were added to each reaction and incubated at 25 °C for 3 hours. The samples were then run on a 1 % agarose gel and the desired bands at 180 bp were cut out and gel purified.

In vitro transcription was carried out using the T7 MegaScript kit (Ambion) with biotin-labelled dUTP (Roche). Equimolar amounts of purified DNA from gBlock1 and gBlock2 were combined and 2 µl 10x buffer, 5.5 µl DNA template (280 ng), 5 µl biotin-UTP (Roche), 1 µl unlabelled rUTP (100 mM), 1.5 µl rATP (100 mM), 1.5 µl rCTP (100 mM), 1.5 µl rGTP (100 mM) and 2 µl T7 enzyme mix were combined. The reaction mixture was incubated at 37 °C overnight.

To remove any remaining template DNA the samples were treated with 1 µl Turbo DNase (Life Technologies) for 15 minutes at 37 °C. The RNA was then purified using the MEGAclear kit (Ambion) following the manufacturer's instructions. The RNA was eluted in 50 µl elution solution. The size and integrity of the RNA were assayed by running 2 µl on a 2 % agarose gel and the final concentration of RNA baits was determined by Nanodrop.

2.10 Generation of biotinylated RNA oligonucleotides for capturing the viral genome from gDNA

An overnight culture from a glycerol stock made from a single colony of plasmid pSP64 HPV16 (Cinzia Scarpini) was prepared with 5 mL LB broth and 100 µg/mL ampicillin and incubated at 37 °C with gentle shaking overnight. Following incubation, plasmid DNA was extracted using QIAprep® Spin Miniprep Kit (Qiagen) according to the manufacturer's instructions. The quantity and purity of the DNA was determined by Nanodrop analysis. To amplify the HPV16 genome, four primer sets were designed across the entire W12E genome. Four PCR reactions were set up, each with a separate primer pair (HPV16_block1-4 for and HPV16_block1-4_rev; Table 2.5); 1 µl plasmid DNA (10 ng/ul), 2 µl dNTP mix (2.5 mM), 10 µl Expand High Fidelity buffer (10x), 0.5 µl forward primer (100 µM), 0.5 µl reverse primer (100 µM), 0.75 µl Expand High Fidelity enzyme mix (Roche) and 82.5 µl dH_2O. A touchdown PCR was run to ensure specific amplification of the DNA. The PCR conditions were the following: 94°C for 5 minutes, followed by 13 cycles of 94°C for 1 minute, 74-62°C for 1 minute (Δ-1°C/cycle), 68°C for 8 minutes, followed by 22 cycles of 94°C for 1 minute, 62°C for 45 seconds, 68°C for 8 minutes, and then 68°C for 10 minutes before cooling to 4°C. PCR products were then run on a 1.5 % agarose gel. T7 promoter sequences were added to one side of the PCR product during the

PCR amplification. Sequences were *in vitro* transcribed using biotin-UTP and purified as described above. An equimolar mix of the four full length (2,000 nt) RNA products was then fragmented to ~150 nt for use in the capture reaction. Chemical fragmentation of the RNA occurred by combining 250 ng of RNA, 100 mM Tris pH 8.0 and 4 mM $MgCl_2$ in 10 µl and incubating at 95°C for 8 minutes in a PCR machine. Five µg RNA were fragmented using this method.

The fragmentation reactions were then pooled into two 100 µl samples and precipitated using 1x volume (100 µl) ice-cold isopropanol and 1/10th volume (20 µl) ammonium acetate Stop Solution (Ambion) and incubated at -20°C for 2 hours. Following incubation, the RNA was spun at 14,000 rpm at 4°C for 20 minutes to pellet the RNA. The pellet was then washed twice with 500 µl 75 % ethanol and resuspended in 20 µl dH2O. The resulting RNA was then quantified using the Nanodrop.

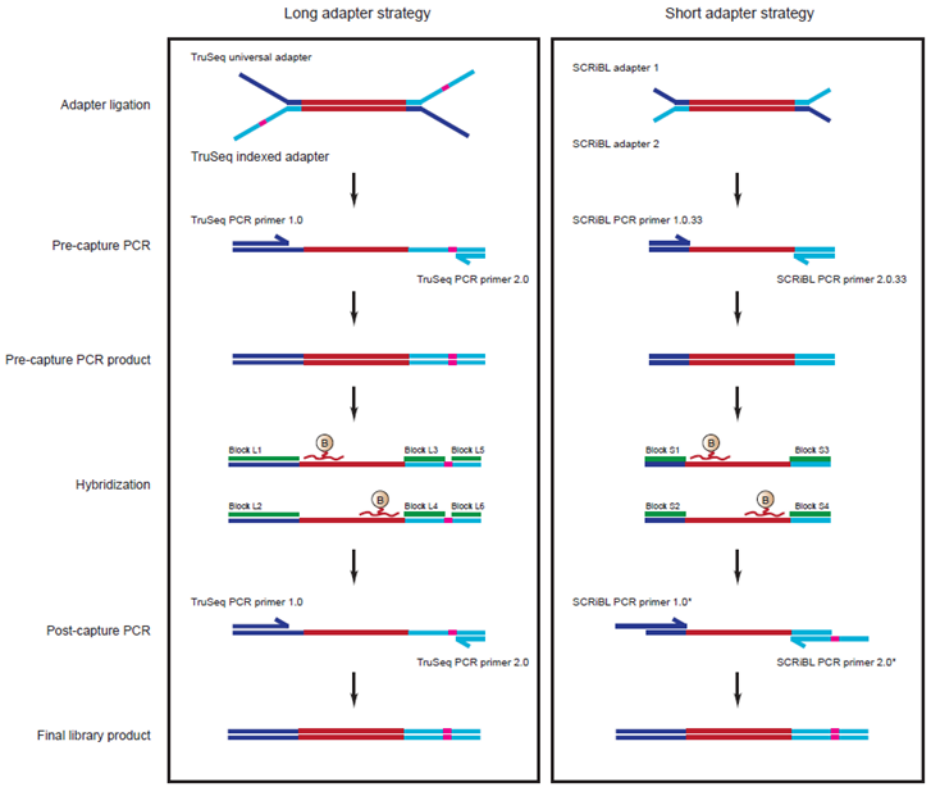


Figure 2.2 | Barcoding and blocking strategy for solution hybridisation capture

Traditionally, long adapter, already containing the desired barcode were ligated during the Hi-C protocol (mCMV Hi-C replicate I). Therefore longer and more blockers needed to be added during the hybridisation step to minimise concatermerisation. In our new strategy, minimal adapters are ligated before the capture step, increasing specificity during the enrichment. Barcodes and flow-cell attachment region were introduced in the post-capture final PCR amplification (human samples and replicate II mCMV Hi-C).

2.11 Solution hybridization capture of libraries with biotin RNA-target baits

The amount of Hi-C library DNA or genomic DNA library captured was determined by the concentration of each library and ranged from 500-2000 ng. To prepare the Hi-C library (pond) for capture with RNA baits, the appropriate volume was transferred into a 1.5 mL LoBind Eppendorf tube, and concentrated using a vacuum concentrator (Savant SPD 2010, Thermo Scientific). After evaporation of all liquid, the Hi-C DNA pellet was resuspended in 5 µl dH_2O and transferred into a fresh LoBind Eppendorf. 2.5 µg mouse cot-1 DNA (Invitrogen) and 2.5 µg sheared salmon sperm DNA (Ambion) were added as blocking agents. To prevent concatemer formation during hybridisation, 1.5 µl blocking mix (300 µM) were added (equimolar mix of four oligo blockers: P5_b1_for_33, P5_b1_rev_33, P7_b2_for, P7_b2_rev) resulting in a 10 µl reaction mixture. This was resuspended thoroughly, transferred into PCR strip tubes (Agilent 410022), closed with a PCR strip tube lid (Agilent optical cap 8x strip) and kept on ice until use. A master mix of the 2.23x hybridisation buffer was then prepared; 167.25 µl 20x SSPE (Gibco, 11.15x final), 66.9 µl 50x Denhardts (Invitrogen, 11.15x final), 6.69 µl, 500 mM EDTA (Gibco, 11.15 mM final), 6.69 µl 10 % SDS (Promega, 0.223 % final) and 52.47 µl H_2O. The hybrisation buffer was mixed thoroughly and heated to 65°C for at least 5 minutes. 30 µl were then aliquoted per capture reaction into a PCR strip, closed with a PCR strip tube lid and kept at room temperature.

Biotinylated RNA baits were used in a ratio of 1:12 to Hi-C libraries (25 ng biotinylated RNA baits per 300 ng Hi-C library) for the human samples and in a ratio 1:2 for the mouse libraries (200 ng biotinylated RNA baits per 400 ng Hi-C library). The corresponding amount of RNA baits were transferred to a LoBind Eppoendorf tube and was made up to a volume of 5.5 µl with H_2O. Subsequently, 30 units (1.5 µl) SUPERase-In (IAmbion, 20 U/µl) were added (7 µl total), mixed, transferred into PCR tubes, closed with a strip tube lid and kept on ice. Biotinylated RNA baits for capture DNA-Seq were used in a ratio of 1:3.33 (300 ng RNA baits per 1,000 ng genomic DNA library); these baits were prepared in the same way as above.

The PCR machine (PTC-200, MJ Research) was set to the following program; 95°C for 5 minutes, 65°C endlessly. The PCR strip containing the pond Hi-C libraries was transferred to the PCR machine in the position marked red (Figure 2.3a), the PCR program was started and the DNA denatured. Once the temperature returned to 65 °C the PCR strip containing the hybridisation buffer was transferred to the PCR machine in the position marked in blue (Figure 2.3b) and incubated for 5 minutes. Following this, the final PCR strip containing the biotinylated RNA bait was transferred to the PCR machine in the position marked in green (Figure 2.3c) and incubated for 2 minutes. After 2 minutes, 13 µl of hybridisation buffer were pipetted into the 7 µl RNA baits (blue into green). This was immediately followed by pipetting 10 µl of the Hi-C library into the hybrisation buffer:RNA mix (red into green). The remaining PCR strip was

closed with a new strip tube lid and the reactions were incubated for 24 hours at 65°C, in a total reaction volume of 30 µl.

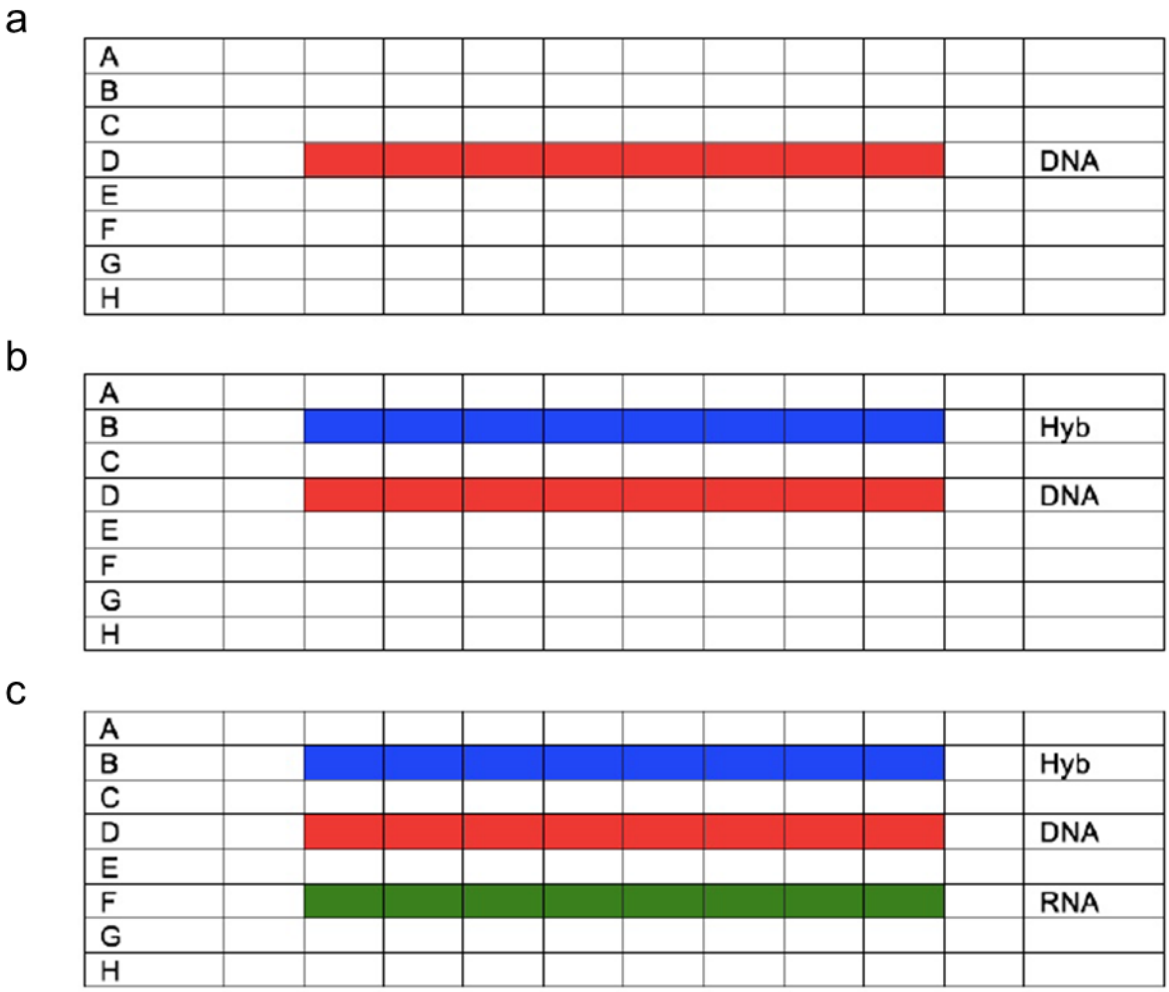


Figure 2.3 | PCR machine set up for the hybridisation of RNA baits to DNA libraries

(a) PCR strip containing DNA (red) is placed in row D at 95°C for 5 minutes. **(b)** Once the temperature has cooled to 65°C the PCR strip containing hybridisation buffer (blue) is put in row B in the machine and incubated for five minutes. **(c)** After the PCR strip containing RNA (green) is added to row F and incubated for 2 minutes before 13 µl hybridisation buffer is added to the RNA (blue into green) immediately followed by the transfer of 10 µl DNA library into the RNA mix (red into green).

Captured DNA/RNA hybrids were enriched using Dynabeads MyOne Streptavidin T1 beads (Life Technologies). 60 µl T1-beads per captured library were aliquoted into a LoBind Eppendorf and washed three times in 200 µl binding buffer (BB:1 M NaCl, 10 mM Tris-HCl pH 7.5, 1 mM EDTA). With the streptavidin beads in 200 µl BB the entire hybridisation reaction from the 24 hour incubation was transferred into the beads and mixed. This was then incubated at room temperature on a rotating wheel. After 30 minutes, the beads were reclaimed on a magnetic separator and the supernatant was discarded. The beads were then resuspended in 500 µl wash buffer I (WBI: 1x SSC, 0.1 % SDS) and incubated at room temperature for 15 minutes with agitation every 2-3 minutes. Following incubation, the beads were once more

reclaimed and the supernatant was discarded. The beads were then resuspended in 500 µl wash buffer II (WBII: 0.1x SSC, 0.1 % SDS) pre-warmed to 65°C, and incubated at 65°C for 10 minutes with agitation every 2-3 minutes. This was repeated for a total of 3 washes in WBII. The beads were reclaimed, the supernatant was discarded, the beads were resuspended in 200 µl 1x NEBuffer 2 and immediately transferred into a fresh LoBind Eppendorf tube. Tubes were placed immediately back on the magnetic rack and the supernatant was removed. Finally the streptavidin beads (with bound captured DNA/RNA) were resuspended in 30 µl 1x NEBuffer 2 and transferred into a fresh tube. To determine the optimal number of PCR cycles for SCRiBL Hi-C library amplification, test PCRs were set up as previously described with PCR cycle numbers tested 9, 12 and 15. The amount of amplified DNA was then checked by running the entire reaction on a 1.5 % agarose gel. In the final reaction primer pairs consisted of one TruSeq adapter reverse compliment and the TruSeq universal adapter (see Table 2.4 for the barcodes used and Table 2.4 for the respective primer sequences). Each primer pair introduced a library-specific barcode. The samples from the complete PCR reaction were pooled and purified using two sequential rounds of SPRI bead selection at 1x and 1.8x volume. The captured libraries were finally eluted from the beads in 20 µl TLE (10 mM Tris pH 8, 0.1mM EDTA). Before sequencing, the quality and quantity of all libraries were checked by Bioanalyzer (Agilent) and Kapa Q quantitative PCR (Kapa Biosystems).

Sequencing for all Hi-C and capture Hi-C libraries, as well as for the captured genomic DNA libraries, was performed on Illumina HiSeq 2500 generating 50 bp paired-end reads (Sequencing Facility, Babraham Institute). CASAVA software (v1.8.2, Illumina) was used to make base calls and reads failing Illumina filters were removed before further analysis.

Table 2.4 | Barcodes introduced to Hi-C and SCRiBL libraries

Hi-C library	**barcode #**	**barcode Sequence**	**SCRiBL library**	**barcode #**	**barcode sequence**
mock II	2	CGATGT	mock II BAC	2	CGATGT
2 hpi II	4	TGACCA	2 hpi II BAC	7	CAGATC
4 hpi II	7	CAGATC	2 hpi II mCMV	4	TGACCA
48 hpi IV	16	CCGTCC	4 hpi II BAC	10	TAGCTT
W12D2 I	6	GCCAAT	4 hpi II mCMV	8	ACTTGA
W12D2 II	12	CTTGTA	48 hpi II BAC	11	GGCTAC
W12G2 I	6	GCCAAT			
W12G2 II	12	CTTGTA	W12D2 I	11	GGCTAC
			W12D2 II	4	TGACCA
gDNA			W12G2 I	16	CCGTCC
W12D2 I	4	TGACCA	W12G2 II	8	ACTTGA

W12G2 II	10	TAGCTT	W12A5 I	10	TAGCTT
W12A5 I	2	CGATGT	W12A5 II	2	CGATGT
W12H II	8	ACTTGA	W12F I	11	GGCTAC
W12F II	1	ATCACG	W12F II	4	TGACCA
			W12H I	16	CCGTCC
			W12h II	8	ACTTGA

Table 2.5 | Primer and adapter sequences

Name	**Sequence 5`-3`**
SCRiBL TruPE adapter 1	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
SCRiBL TruPE adapter 2	ACACTCTTTCCCTACACGACGCTCTTCCGATC*T
PE_PCR_1.0.33	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE_PCR_2.0.33	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq universal primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
TruSeq_rc_2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq_rc_4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq_rc_6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq_rc_7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq_rc_8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq_rc_10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq_rc_11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq_rc_12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
TruSeq_rc_16	CAAGCAGAAGACGGCATACGAGATCGGGACGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
Lou T7 BamHI for:	TCTAGTCGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAG
Lou T7 BamHI rev	[Phos]GATCCTCGCCCTATAGTGAGTCGTATTACAATTCACTGGCCGTCGACTAGA[SpcC3]
Lou T7 BglII:	TCTAGTCGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGA
Lou T7 BglII rev:	[Phos]GATCTCGCCCTATAGTGAGTCGTATTACAATTCACTGGCCGTCGACTAGA[SpcC3]
HPV_block_1_for	AATTGTAATACGACTCACTATAGGGAGACCCATGTACCAATGTTGCAG
HPV_block_1_rev	ATCCCGAAAAGCAAAGTCAT
HPV_block_2_for	AATTGTAATACGACTCACTATAGGGAGATGACTTTGCTTTTCGGGATT
HPV_block_2_rev	TTGCTTCCAATCACCTCCAT
HPV_block_3_for	AATTGTAATACGACTCACTATAGGGAGAAGATGTGATAGGGTAGATGATGGA
HPV_block_3_rev	TGTAATTAAAAAGCGTGCATGTG
HPV_block_4_for	AATTGTAATACGACTCACTATAGGGAGACATACACATGCACGCTTTTT
HPV_block_4_rev	TTCCCCTATAGGTGGTTTGC
RPL13A_B_for	AGGCGTGTTACTGGAAGTCG
RPL13A_G_for	GAGCCTTGCTGGTCTTCGTT
RPL13A_D2J_for	GGTGCGATCGATCCTCATGAAA
RPL13A_D2J_rev	GTGACTGACAGCTGGGCATA
P5b1for33	ACACTCTTTCCCTACACGACGCTCTTCCGATCdd

P5b1rev33	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
P7b2for	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCdd
P7b2rev	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Calr_1_for	AGGGAGAAAGGGGATGAGAA
Calr_2_for	CGGACCATTTCAGAACACCT
h4i-2-for	GCCACTTCCCTTCAGTCAAA
h4f-1-rev	CTTCCGCAGGTTCCCTAAGT

2.12 General processing and analysis of genomic data

2.12.1 Genomic features

Overlaps, distance calculations and filtering for genomic annotations, peaks or other features was performed in Seqmonk (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/), or by manipulation in the statistical language R, making additional use of the GenomicRanges package (https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html) from the BioConductor webpage. Gene annotations were taken from the core gene annotation for GRCm38/mm10 or GRCh37/hg19 provided by Seqmonk, which is based on the Ensembl gene annotations. For analysis of Hi-C and SCRiBL data, the respective reference genome was divided into fragments based on BglII or MboI restriction enzyme recognition sites, respectively, depending on the enzyme used for library generation. This was done using the hicup_digester perl script, which is part of the HiCUP tool (Wingett et al., 2015).

2.12.2 Pre-processing and alignment of sequencing data

Data-set specific details are provided below, but in general, sequenced reads were trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove low quality base and adapter sequences, and basic quality control checks were performed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were aligned to the mm10 or the hg19 with either the mCMV genome or the HPV16 genome as an additional chromosome (details given below).

2.12.3 Read counts and visualisation

BAM files were imported into Seqmonk for visualisation and initial analysis of raw data. Seqmonk was used to generate raw and normalised read counts by quantitating reads overlapping defined regions (termed “probes” in Seqmonk). Screen shots showing comparisons between datasets show the average signal of two biological replicates, normalised by total read count. Read distributions around specific features were calculated in Seqmonk, using the “probe trend plot” tool, to calculate either relative or cumulative distribution plots.

2.13 Analysis of 4sU-Seq data

2.13.1 Read processing and read counts

4sU-Seq reads were trimmed and aligned with Tophat (Trapnell et al., 2009), using default parameters. Bam files were additionally filtered for pairs aligning to ribosomal DNA sequences and those were removed (Caroline Friedel). For calculation of feature coverage and strand specificity, mapped reads were imported into Seqmonk using default settings for paired-end RNA-Seq data.

2.13.2 Gene expression categories

Promoter expression levels were determined from non-infected NIH-3T3 data using counts of reads per million per kilobase of exon (RPKM) to correct for transcript length. Genes with a value of less than 0.7 RPKM were placed into a single category of non-expressed genes. The remaining genes were divided into four equally sized quartiles of expressed genes. For defining expressed genes for soft clustering, a gene had to have a greater than 0.7 RPKM value in both replicates, in at least one of the nine time points of infection.

2.13.3 Fuzzy c-means clustering of newly transcribed RNA

RPKM values of expressed genes were averaged between replicates for genes at the individual time points, in order to combine replicates. Expression values of combined expressed gens were standardized to have a mean value of zero and a standard deviation of one. This was done by using the “standardize” function of the “Mfuzz” R package (Kumar & M, 2007). The fuzzifier *m* was determined by using a direct estimate proposed by (Schwammle & Jensen, 2010) and encoded by the Mfuzz function “mestimate”. The optimal cluster number *c* was determined in two ways: by monitoring the minimum Euclidian distance D_{min} across a range of *c*, commonly known as “elbow method” and by using a function in R called „clustergram“ (https://raw.githubusercontent.com/talgalili/R-code-snippets/master/clustergram.r). For the elbow method“, I made use of the Mfuzz package function „Dmin“, using the *m* previously estimated by the „mestimate“ function. The „clustergram“ function was downloaded and run to monitor cluster separation from 2 to 25 clusters for a total of 7 times. Only stable optimal cluster numbers from both approaches were carried forward. The actual soft clustering was done by utilising the „mfuzz“ function with the estimated optimal *c* and *m* and was visualised using the „mfuzz.plot2“ function. Cores of the clusters were isolated using the „acore“ function.

2.14 Analysis of ATAC-Seq data

2.14.1 Initial data processing

One intrinsic ATAC-Seq problem is that many DNA fragments are much shorter than the number of sequencing cycles. Thus, many reads also contain the 19 bp Tn5 mosaic sequence

inserted by tagmentation, and potentially prefixes of the sequencing adapters. In these cases, the paired-end reads have a specific layout:

Read 1: [DNA fragment] [mosaic][5'-adapter]
Read 2: [DNA fragment (reverse complementary)] [mosaic] [3'-adapter]

An in-house program allowing for up to seven mismatches prior to read mapping checked this layout, and if detected, mosaic and adapter sequences were trimmed. Then we used Bowtie 2 (version 2.1.0) to map all reads against the mouse (mm10) and mCMV (NC_004065.1) genomes allowing for a maximum fragment length of 2000 (-X 2000) and leaving all other parameters set to default. The resulting SAM files contain the coordinates of both reads for each pair. To facilitate subsequent analyses, we generated a BAM file containing one entry for each mapped read pair by computing the mapping coordinates of the original DNA fragment from the coordinates of the two reads. Blacklisted mm10 regions (https://sites.google.com/site/anshulkundaje/projects/blacklists) were excluded from all downstream analysis.

2.14.2 Accessibility peaks and read counts

ATAC-Seq BAM files were used to call peaks using MACS2 (Zhang et al., 2008), using the following parameters: -f BAM -g mm -q 0.01 --nolambda --nomodel --keep-dup all. Shared peak locations were defined in R using the “GenomicRanges” package, with a minimum overlap of 1 bp. Aligned BAM files were imported into Seqmonk as single ended data and the 5` end positions of reads were extracted. These single base-pair reads were used to generate read counts for specified regions, including cumulative count profiles of read distributions around specific features. All quantitation was done keeping duplicates.

2.15 Analysis of Hi-C and SCRiBL data

2.15.1 Initial data processing

Hi-C and SCRiBL sequencing reads were processed using HiCUP (Wingett et al., 2015), which truncates reads containing putative Hi-C ligation junctions to improve mapping efficiency. HiCUP aligns reads in each pair independently using Bowtie 2 (Langmead & Salzberg, 2012) to map interacting loci. HiCUP additionally filters out experimental artifacts, such as circularized reads, re-ligation and non-ligated fragments. Furthermore, duplicated reads and di-tags not falling within 120-800 bp were removed. For SCRiBL libraries, the resulting Bam files were additionally filtered for read-pairs where at least one end mapped to a captured restriction fragment end, utilising a custom perl script written by Steven Wingett.

2.15.2 Virtual 4C profiles and interaction counts

Aligned Bam files were imported into Seqmonk using the Hi-C data option and keeping only those reads with a mapping quality score > 20, while all other parameters were left as default. Virtual 4C Hi-C profiles for visualization and analysis were generated using the “Hi-C other ends” tool, using either a single BglII fragment or the entire viral genome (both HPV and mCMV).

SCRiBL of the HPV16 genome approximates multiplexed 4C experiments with multiple viewpoints. However, due to the limited number of captured di-tags per HPV 16 MboI fragment, the entire provirus was treated as a single viewpoint. The tag counts per HPV16 interacting fragment was determined and these counts were supplied to the Bioconductor package FourCSeq (Klein et al., 2015). To find significant interaction between the viewpoint and the fragments, FourCSeq applies a variance stabilizing transformation to the counts and calculates a distance-dependent monotone fit. Z-scores are derived from the fit and agreement between replicates is taken into account. Significant fragments were those with z-scores greater than 2 in both replicates and an adjusted p-value of less than or equal to 0.05 in at least one replicate.

2.15.3 Significant interactions with GOTHiC

Hi-C and SCRiBL BAM files were converted to a format compatible with the BioConductor package GOTHiC (Mifsud et al., 2017). To find significant interaction between distal locations GOTHiC implements a cumulative binomial test based on read depth. This was used to identify regions of the human or the mouse genome in contact with the HPV or the mCMV pseudo-chromosome, at a resolution of 1 kb for the human data and a resolution of 200 kb for the mouse data. Significant di-tags between the human genome and the HPV16 genome were visualized using Circos (Krzywinski et al., 2009). The BAM outputs were converted to BED format and modified to be compatible with the circular visualization tool *Circos*. The HPV16 genome was split into 500 bp bins and the count per bin determined from chimeric human-HPV16 di-tags. The counts, HPV16 MboI restriction map and gene coordinates are annotated on the plot.

2.15.4 Heatmap generation

To produce heatmaps, the genome was divided into equal sized loci of varying bin size and each interaction was binned according to the location of both ends to produce the matrix. Compressed contact matrices were generated by either Juicer (Durand et al., 2016b) or HOMER at varying resolutions and were visualized with Juciebox (HPV16 Hi-C) (Durand et al., 2016a) or Java TreeView (mCMV Hi-C). Heatmaps were normalized by depicting the observed to expected interactions, assuming each region has an equal chance of interacting with every other region in the genome and that regions are expected to interact depending on their linear

distance along the chromosome. As part of the heatmap generation in HOMER, an output file with 10 kb resolution is generated providing the genome-wide distance decay, which was plotted in R for the individual time points. Furthermore, by providing a file of active and inactive regions of the genome (determined by interaction with the nuclear lamina in MEFs (Peric-Hupkes et al., 2010)) the distance decay for those regions can be determined. Contact matrix information can also be modified by HOMER to be directly fed into *Circos* with the –circos option.

2.15.5 A/B compartmentalization, TAD calling and insulation score calculation

Principal component analysis (PCA) was described to partition the genome into two compartments, A and B (Lieberman-Aiden et al., 2009). Csilla Varnai applied PCA to normalized interaction matrices of individual chromosomes at a resolution of 200 kb, by utilizing the runHiCpac.pl script provided by HOMER. By providing a seed region of H3K4me3 (active promoters in MEFs) HOMER determines if the resulting values need to be multiplied with -1 in order for the A compartment to have positive values. Changes in compartments were calculated in R; regions had to change significantly from -5 to +5 or *vice versa*.

TADs were calculated using the directionality index as previously described (Dixon et al., 2012) and implemented in the "findHiCDomains.pl" provided by HOMER. This was done with the following settings: minIndex 0.7, minDelta 1.5, res 10000, superRes 50000, window 50000 and maxError 1. TAD boundary overlap was calculated in R using the GenomicRanges package, allowing shifts of 1 bin to either side. Boundary strength was calculated by making the ratio of internal/external reads per TAD. Variation of the insulation of all TADs was calculated by assuming the true mean to be 0 (no differences between infected and non-infected cells). The mean and the difference between each insulation score was calculated for each TAD. Data were ordered by means and the standard deviation of the differences was calculated using the 50 points ahead and behind it. The" pnorm" function in R was used to calculate p-values. TADs with a p-value < 0.05 were termed closing TADs and the remaining ones were termed non-responsive. Overlap of each category with genomic features was calculated in R.

To measure the topological domain structure along the human chromosome 5 in the in the HPV16 infected clones D2 and G2, we computed an average insulation score profile at the TAD boundaries around the integration sites. The insulation score is the standardized -log enrichment of contacts between the downstream and upstream 300 kb regions (-log (a/(a+b1+b2)) where a is the number of contacts between, and b1 and b2 the number of contacts within the upstream and downstream 300kb regions). Using this definition, a more positive insulation score indicates a stronger TAD boundary.

2.15.6 Open chromatin index (OCI) calculation

OCI values were calculated using SeqMonk. Rolling windows were created across the genome (200 kb unless otherwise stated). Raw Hi-C di-tag counts were calculated for each window; those with no counts or extremely high counts were removed. The “cis/trans” quantitation method was then used on the remaining bins. The chromosomal median values were subtracted for all windows. The “smoothing subtraction quantitation” using a window size of 20 Mb was used to normalise for biases along the lengths of the chromosomes.

2.16 Gene ontology and TFBS prediction

Gene ontology analysis of “ Biological Processes” was done using the DAVID bioinformatics analysis suite (Huang da et al., 2009). *In silico* TFBS prediction was done using HOMER (http://homer.ucsd.edu/homer/) with default settings, with either -500 bp to +100 bp around TSS or defined ATAC-Seq peaks around TSS.

2.17 General statistics and data visualization

Unless otherwise specified, graphs were produced in R and Seqmonk. Box plots in all cases show the median, interquartile range and range; whilst outliers, defined as values > 1.5 times the interquartile range away from the box are not plotted. Bean plots were generated using the beanplot R package (https://cran.r-project.org/web/packages/beanplot/index.html). Scatter plots were visualized using ggplot. Venn diagrams were produced using the VennDiagram R package (https://cran.r-project.org/web/packages/VennDiagram/index.html). Non-Hi-C heatmaps were generated using the heatmap.2 function in R. Some barcharts and line graphs were plotted with Microsoft Excel. Calculation for Pearson`s correlations coefficient, Fisher`s exact test and the Wilcoxon rank-sum test were performed in R (see individual figure legends for details). Figures were assembled in Adobe Illustrator CC, Inkscape and Microsoft PowerPoint.

3 Transcriptional changes upon lytic mCMV infection

3.1 Introduction

Like all other herpesviruses, mCMV has co-evolved with its respective host over millions of years, which has provided sufficient time for the viruses to master host-cell modulation to facilitate their own needs and survival. Host gene expression is subjected to strong and rapid alterations during herpes viral infections, induced by viral and antiviral mechanisms. Herpes viruses have large genomes and encode for a huge number of different genes, of which most are dispensable for viral replication in cell culture. Almost certainly, their products are required for immunomodulation of the host cell. Understanding how hosts and pathogens modulate gene expression during the host-pathogen interaction is key to uncover the molecular mechanisms that regulate disease progression.

Several high-throughput studies have addressed the transcriptional response of cells to lytic hCMV infection by analyzing temporal changes in total RNA levels (Bresnahan & Shenk, 2000; Browne et al., 2001; Challacombe et al., 2004; Hertel & Mocarski, 2004; Sarcinella et al., 2004; Terhune et al., 2004). These studies revealed that lytic hCMV infection altered the expression of numerous host genes involved in a variety of processes including inflammation, innate immunity, cell cycle progression, cellular metabolism and cell adhesion. Measurements of total RNA however mask the contribution of *de novo* RNA synthesis of the host and the virus, especially early upon infection by the large amounts of virion-associated RNAs, which are nonspecifically incorporated by the viral particles and are delivered to the newly infected cells. To overcome this hurdle and to study changes in transcriptional synthesis rates in mouse cells upon lytic CMV infection, a study has employed metabolic labeling of newly transcribed RNA with 4sU on lytically infected NIH-3T3 fibroblasts (Marcinowski et al., 2012). This provided a unique opportunity to study viral RNA synthesis without the interference of virion-associated and stable host RNAs and revealed three major findings about viral gene expression. 1) All three classes of viral gene transcripts (immediate early, early and late genes) showed a prominent peak at 1-2 hpi at high and low MOI. 2) A rapid and strong suppression of all three classes of viral genes by 5-6 hpi. 3) Very constant levels of gene transcription or even increasing suppression of some viral genes at later stages of infection, despite the dramatic increase in viral copy number due to extensive viral replication. Further, this labeling technique improved the detection of gene clusters with different expression kinetics as well as downstream *in silico* promoter analysis compared to total RNA measurements. Thus, this approach provides an ideal mean to gain insights into the molecular mechanisms and the TFs involved in host gene modulation upon infection.

Gene activation by sequence-specific DNA-binding TFs requires the removal of nucleosomes from the target sites to allow those factors to access and bind target sequences. Thus, the

removal of nucleosomes from TSSs and regulatory regions has been proposed as a key mechanism of gene activation (Mavrich et al., 2008). Alternative mechanisms such as altering the accessibility of DNA on the surface of the nucleosome by ATP-dependent chromatin remodelers, the incorporation of certain histone variants or destabilising nucleosome contacts by acetylation of lysine residues have been proposed in different contexts (Sexton et al., 2014; Soufi et al., 2015). Within accessible regions, direct interactions between proteins, such as TFs, and DNA prevent enzymatic cleavage by DnaseI or Tn5 transposase, resulting in protected "footprints" of these proteins (Brenowitz et al., 2001). These footprints allow genome-wide profiling of TF binding in an unbiased manner and allow constructing gene regulatory networks (Sullivan et al., 2014).

Reversely, a model in which the chromatinisation of the herpes simplex virus type 1 (HSV-1) genome is a cellular defense mechanism to silence the incoming viral genomes has been proposed (Gibeault & Conn, 2016). The host defense is believed to be counteracted by the virus by expressing proteins that prevent or disrupt the stable chromatinisation of HSV-1 genomes, such as VP16, ICP0 and ICP4. This also holds true to some extend for the hCMV immediate early protein 1 (IE1) (Zalckvar et al., 2013). The CMV genome does not carry any nucleosomes in the viral particles, but has been proposed to harbor nucleosomes at non-random positions once located in the host cell nucleus. Early in infection, histones bind to the viral DNA in a sequence specific manner; whereas late in infection nucleosomes redistribute extensively to establish patterns mostly determined by the viral IE1 protein. These temporal nucleosome occupancy differences correlate inversely with changes in nascent viral gene transcription. In the absence of IE1 the hCMV genome is much less dynamic in terms of nucleosome positioning compared to the wild type virus (Zalckvar et al., 2013).

3.2 Objectives and outline

In this chapter, I will provide the most comprehensive description to date on the transcriptional changes that are occurring during lytic mCMV infection, within the host and the virus. Fuzzy c-means clustering will group host genes based on their expression pattern into functional clusters. *In silico* promoter analysis will provide first useful insights into which TFs might be involved in regulating the host gene expression and at which time point of infection they start to play a role. By integrating ATAC-Seq data with transcriptional activity, I will corroborate the importance of nucleosome positioning on transcriptional activity. ATAC-Seq can further be utilised to improve the definition of proximal promoter regions (PPRs), which then will be used to improved TFBS prediction by reducing the background.

Additionally, information on the nucleosome positioning and on the function of mCMV chromatin during infection is limited, in part because no genome-wide studies are available.

For instance, it is not known whether the viral DNA forms nucleosomes in an organized fashion, and how nucleosome occupancy may regulate the cascade of viral transcription.

To address these questions, I generated high-resolution spatial and temporal maps of nucleosome occupancy and measured nascent transcription across the entire mCMV genome and the host genome after infection.

3.3 Results

3.3.1 Measuring nascent transcription using opposing strand specific 4sU-Seq

To determine changes occurring in the mouse transcriptome upon lytic mCMV infection, I performed nascent RNA sequencing analysis by employing metabolic labeling of newly transcribed RNA with 4sU for an hour at multiple time points during infection. To this end, NIH-3T3 mouse fibroblasts, which are permissive to mCMV infection, were infected with BAC-derived mCMV Smith strain at an MOI of 10 (as described in 2.1.3) in two biological replicates per sample and newly transcribed RNA was labeled for an hour (as described in 2.4). In order to achieve robust quantification for the reference time point (mock-infected samples) and to determine the inherent technical noise of 4sU experiments in general, I generated two technical replicates per biological replicate for the uninfected mock-treated cells, resulting in four libraries from non-infected NIH-3T3. A previous study reported the onset of viral genome replication occurring at ~15 hpi, determined by qPCR on M54 (viral DNA polymerase) and southern blot analysis of concatemeric viral DNA (Marcinowski et al., 2012). I, therefore, decided to label newly transcribed RNA prior to the onset of viral DNA replication, for an hour at each (0-1 hpi, 1-2 hpi, 3-4 hpi, 5-6 hpi, 11-12 hpi), just after the onset of viral gene DNA replication (17-18 hpi), when first infectious particles are released (24-25 hpi) and at a late stage of infection (47-48 hpi). Prior to library preparation and sequencing, enrichment for newly transcribed RNA was tested and confirmed by qPCR (supplementary figure 1). Opposing strand-specific RNA-Seq libraries were generated by BGI Tech Solutions in Hong Kong and were sequenced on the Illumina Xten machine with a 100 bp paired-end output. Reads were reversed in their orientation for analysis, to match the strand they originated from. Reads mapping to the viral genome were discarded while looking for changes in the host transcriptome and vice versa. A representative locus of the host genome is displayed in Figure 3.1a. Clear enrichment of reads overlapping annotated genes, especially exons, could be observed for the given locus. The presence of introns in my data suggests successful enrichment of newly transcribed RNA. Furthermore, a strong signal of anti-sense transcription at the promoter is detectible for Cops7a and Zfp284.

The obtained sequencing depth ranged between 17-30 million reads per sample uniquely mapping to the host genome (Table 3.1). Based on recommendations given by the ENCODE consortium (https://genome.ucsc.edu/ENCODE), the achieved host transcriptome sequencing

depth is high enough to reliably detect and measure rare, yet physiologically relevant, RNA species (those with abundances between 1-10 copies per cell). The sensitivity of RNA-Seq experiments, including newly transcribed RNA libraries, is a function of the overall coverage, the molar concentration and the transcript length. Hence, I annotated reads to known transcripts (ensemble.org) and quantified their levels in reads per kilobase of exon per million mapped reads (RPKM). This measure facilitates transparent comparison of transcript levels both within and between samples and replicates.

To first assess the quality and reproducibility of the generated results, I compared RPKM values between technical and biological replicates in uninfected mock-treated samples. Both pairs of technical replicates were highly similar (R^2 = 0.99 and R^2 = 0.98, respectively) but reproducibility was also high between biological replicates (Figure 3.1b), with correlation scores, R^2, ranging from 0.96 to 0.97 (Figure 3.1c). Furthermore, the overall high correlation between biological replicates holds true for samples early (3-4 hpi) and late (47-48 hpi) in infection. Summing the replicates over the entire host transcriptome for all independent time points of infection (Figure 3.1d) showed that between ~50-62 % of all reads are falling onto ~65 % of all genes. Around 65-70 % of those reads are falling onto exons. This means that 30-35 % of the reads in my data overlapping genes are falling onto introns and that of all reads between 30-40 % are falling onto exons, even though exons comprise less than 2 % of the genome. Overall, less than 4 % of all reads overlapping genes were in the anti-sense direction after reversing their orientation, indicating successful opposing strand-specific RNA-Seq library generation of newly transcribed RNA. During the mapping process, reads overlapping rRNA were discarded and therefore are not represented in the data. 5-10 % of all reads mapped to the mitochondrial genome.

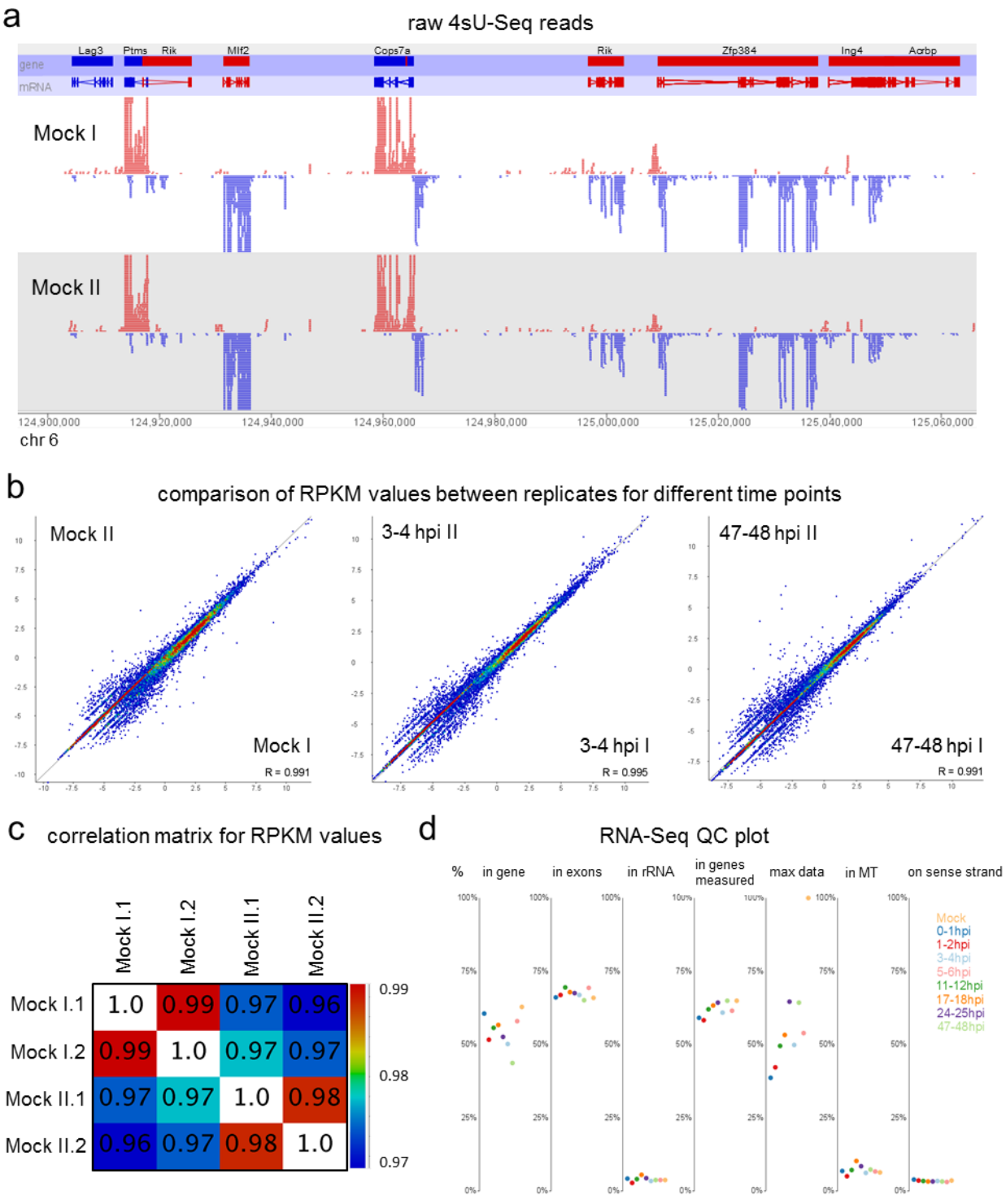


Figure 3.1 | 4sU-Seq quality controls

NIH-3T3 were infected with BAC derived mCMV Smith strain at an MOI of 10, newly transcribed RNA was labelled for an hour at multiple crucial time points during infection (as described in 2.4) and opposing strand-specific libraries were generated and sequenced. **(a)** Raw reads (not reversed in orientation) are depicted for a representative locus of the mouse genome in non-infected cells for both replicates. **(b)** Scatterplots comparing RPKM values of transcripts between replicates for the three indicated time points (mock-infected cells on the left, 3-4 hpi in the middle and 47-48 hpi on the right). **(c)** Heatmap of all obtained pair-wise Pearson`s correlations for all four libraries from non-infected NIH-3T3 calculated for RPKM values of transcripts. **(d)** RNA-Seq quality control plot showing the distribution of all reads from combined replicates for all time points over the given genomic features.

Table 3.1 | 4sU-Seq read numbers

Individual host and viral raw mapped read numbers separated by orientation

sample	host total	host forward	host reverse	virus total	virus forward	virus reverse
Mock I.1	29,109,510	10,066,229	19,043,281	6,559	4,585	1,974
Mock I.2	28,988,695	9,163,052	19,825,643	5,974	4,138	1,836
Mock II.1	25,933,221	8,936,775	16,996,646	6,044	4,108	1,936
Mock II.2	23,668,216	8,098,340	15,569,876	7,253	4,950	2,303
hpi 0-1 I	20,127,790	6,851,784	13,275,106	716,187	497,632	218,555
hpi 0-1 II	20,359,661	6,640,460	13,719,201	454,913	297,813	157,100
hpi 1-2 I	17,587,304	5,817,280	11,770,024	3,435,328	2,323,612	1,111,716
hpi 1-2 II	20,594,876	6,865,896	11,728,980	3,554,701	2,397,326	1,157,375
hpi 3-4 I	22,533,068	7,990,341	14,542,727	5,523,214	3,961,194	1,562,020
hpi 3-4 II	19,826,436	6,758,086	13,068,350	5,431,862	3,758,874	1,672,988
hpi 5-6 I	28,880,627	10,416,431	18,464,196	4,248,317	3,179,565	1,068,752
hpi 5-6 II	22,409,647	8,090,974	14,418,673	2,976,998	2,209,079	767,919
hpi 11-12 I	30,804,295	10,331,274	20,473,021	2,526,920	1,945,296	581,624
hpi 11-12 II	18,707,153	6,434,441	12,272,712	1,630,098	1,247,560	382,538
hpi 17-18 I	21,550,005	7,755,683	13,794,322	4,780,050	3,271,397	1,508,653
hpi 17-18 II	25,008,925	9,003,927	16,004,998	5,939,395	4,047,442	1,891,953
hpi 24-25 I	26,700,069	10,048,037	16,652,032	8,298,217	5,851,025	2,447,192
hpi 24-25 II	25,328,483	9,415,328	15,913,155	8,737,532	5,948,949	2,788,583
hpi 47-48 I	19,644,980	7,372,379	12,272,601	11,421,327	8,029,416	3,391,911
hpi 47-48 II	23,292,665	8,418,877	14,873,788	12,817,013	8,900,591	3,916,422

3.3.2 Soft clustering of differentially expressed genes upon lytic mCMV infection

Having established, that my 4sU-Seq data are of high quality, I wanted to identify genes that display differential nascent RNA levels upon lytic mCMV infection. Marcinowski et al. used a clustering approach based on two-fold differences in abundance at different early time points early in infection (Marcinowski et al., 2012). In this thesis, I extended the number of time points to look at, up into the very late stages of infection, resulting in nine time points to cluster, including the non-infected samples. Only genes that could reliably be detected were subjected to downstream analysis. To pass this cut-off, a gene had to have a RPKM value greater than 0.7 in at least one of the time points in both biological replicates. Accordingly, an absolute number of reads larger than 12 were falling onto that transcript per kb. This resulted in 12,368 detectable transcripts. Figure 3.2a shows the general trend of transcriptional regulation of expressed genes at the individual time points in fold changes compared to the non-infected samples. It is apparent, that as early as 3-4 hpi, the dominant direction of gene regulation is upwards and this trend is increasing in number of regulated genes and in strength with the ongoing infection. Although, around 7-10 % of all expressed genes show a strong (>2 fold) down regulation at each of the individual time points, detectable as early as 2-3 hpi.

Clustering the huge number of genes across the large number of time points exceeds the capabilities of the human brain. Hence, I decided to utilise an unsupervised clustering approach. There are two fundamentally different methods for unsupervised gene clustering, hierarchical clustering and partitioning. In hierarchical clustering, each cluster is subdivided into smaller clusters, forming a tree-shaped data structure or dendrogram, which is shown for my data in Figure 3.2b. On the other hand, partitioning methods, such as k-means clustering or self-organising maps (SOMs), subdivide the data into a predetermined number of clusters, without any implied hierarchy. There is some evidence, that *k*-means and SOM outperform hierarchical clustering (Datta & Datta, 2006). Henceforth, I decided to utilise the k-means partitioning method. Furthermore, there are methods available to estimate the numbers of cluster, best representing the data, and SOMs only seem to outperform k-means clustering for large numbers of clusters (Gat-Viks et al., 2003).

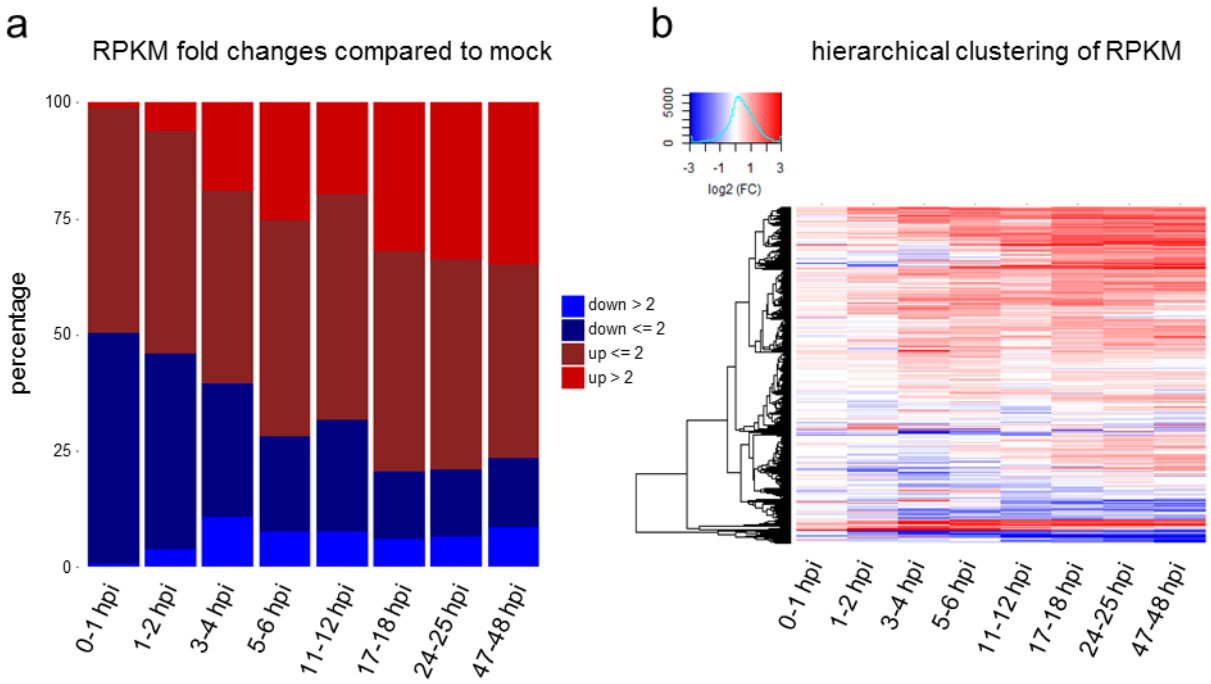


Figure 3.2 | General trends in host gene expression alterations

(a) Fold changes of averaged RPKM values between both replicates at the indicated time points compared to non-infected cells. Coloured for greater or smaller 2 fold up or down-regulation as indicated in the figure. **(b)** Heatmaps including dendrogram obtained from hierarchical clustering of RPKM values of transcripts from replicate 1.

Hard clustering, i.e. assigning a gene to one cluster only is preferable to use when clusters are well defined and separated. This however is barely the case for temporal expression data, where clusters often overlap. Since hard clustering will always fully assign genes to one of the clusters, even though this might not be biological meaningful, it is highly sensitive to noise, such as biological variation arising from variability of the samples and processing noise. To overcome the limitations of hard clustering, I employed soft clustering based on the fuzzy

c-means algorithm first described by Dunn (Dunn, 1973) and improved by Bezdek (Bezdek, 1981), implemented in the open-source statistical language R package called “Mfuzz” (Futschik & Carlisle, 2005). Soft clustering is more noise robust and a priori pre-filtering of genes, which would result in loss of information, can be avoided.

3.3.2.1 Clustering parameter estimation

At first, the “Mfuzz” R package takes RPKM values as an input, but has no in-built function to handle replicates. I therefore decided to merge replicates by averaging the RPKM values from the two replicates for any given expressed gene at all given time points. Normalisation of merged expression values by transformation into RPKM values makes different samples comparable, but since the clustering is performed in the Euclidian space, expression values were further standardised to allow for the comparison of transcripts within a sample. Two parameters need to be determined prior to clustering, the fuzzifier *m* and the number of cluster *c*. By choosing *m*, fuzzy clustering can be tuned so that random data is not clustered, which is a clear advantage over hard clustering methods, such as *k*-means. To achieve this, I utilised a direct estimate, which computes *m* based on the size of the data (Schwammle et al., 2010), resulting in m = 1.355694. A known caveat of all clustering approaches is the challenging determination of the optimal number of clusters. There is no single best criterion for obtaining a partition because no precise and practical definition of “cluster” exists. Here, I used two different methods to reliably and most accurately estimate a useful cluster number *c*.

Firstly, I monitored the minimum Euclidian distance D_{min} between all clusters across a range of *c*, commonly known as “elbow method”. One would expect, that D_{min} declines slower after reaching an optimal *c*. Plotting D_{min} over *c* (Figure 3.3a) vaguely resembles an arm with the elbow being at c = 22. Additionally, cluster numbers of c = 4 and 14 stand out, due to the differences in D_{min} to follow. As a second estimate, I utilised a function called “clustergram” implemented in the statistical environment R. The function calculates weighted PCA means of the cluster centres, orders them according to their respective clusters first component and plots them against *c*. A few interesting observations can be made from the resulting graph (Figure 3.3b).

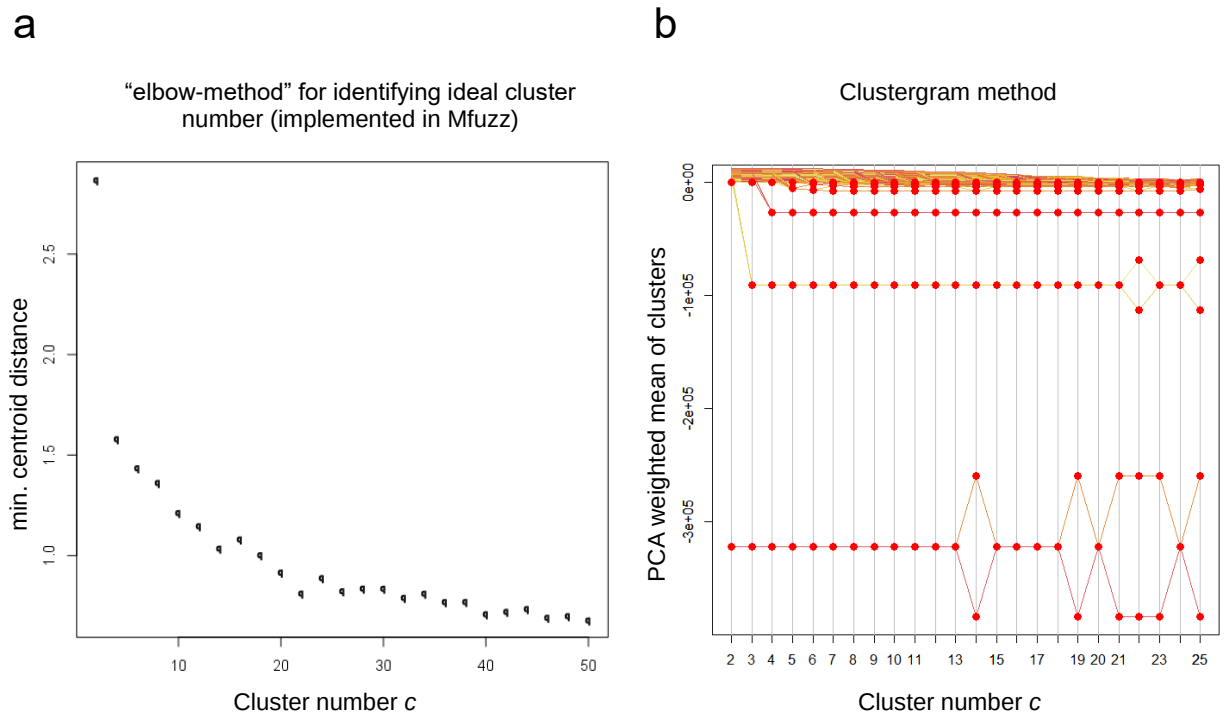


Figure 3.3 | Fuzzy c-means cluster number estimation

Estimation of the cluster number best representing the data was conducted using two independent methods to obtain the most robust result. **(a)** Plotting the minimum distance D_{min} in the Euclidian space between all clusters over the cluster number *c*, vaguely resembles an arm with the elbow being detected at *c*= 22. **(b)** The clustergram R package produces a first principal component weighted mean of the cluster centres from each of the iterations. All data points are then ordered according to their clusters first principal component and plotted against the number of clusters. The shown results were reproducible for 6 separate iteration (supplementary figure 2).

The first few *n* splits, up to *c* = 4 clusters, result in the separation into *c* = n + 1 well spread clusters. Adding the following 9 clusters does not improve the description of the data and we are practically left with the initial four clusters. The smallest cluster number *c* with the best spread of cluster distances is *c* = 14. This estimate was stable when running the algorithm several times (supplementary figure 2). Based on the results obtained from both cluster number estimation methods, I went on and performed the clustering with *c* = 4, 14 and 22 clusters. The most reliable quality measure of a clustering method is how well it actually performs the task at hand. In my case, the hypothesis was that genes from the same cluster will belong to functional clusters i.e. have similar cellular functions (GO terms) and therefore might contain function-specific *cis*-regulatory sequences (TFBS) in their promoter regions. Gene ontology analysis was performed using the David GO online tool. *In silico* TFBS prediction was performed using HOMER (http://homer.ucsd.edu), as described in 2.16.

The Mfuzz package computes gradual membership values of a gene between 0 and 1 indicating the degree of membership of this gene to a given cluster. Thus, effectively reflecting the strength of a gene`s association with a cluster, enabling the definition of a cluster core

defined by tightly co-expressed genes. By setting a threshold α, the inner structure of a cluster can be assessed. The cut-off α was initially chosen manually to give sensible numbers of genes (100-300) for each of the individual clusters, ranging from α = 0.7 over α = 0.8 to α = 0.99 for *c* = 22, 14 and 4 clusters, respectively.

Clustering into four clusters resulted in no cluster significantly associated with a biological process. The obtained resolution was not high enough to separate genes into clusters with defined and significantly associated functions. When clustering into *c* = 22 clusters, some clusters were significantly enriched for specific biological processes, but most of the clusters did not have any significantly associated function. Strikingly, when using *c* = 14 clusters all but one cluster showed functional enrichment. Hence, clusters generated using *c* = 14 clusters were subjected to α-core improvement and further downstream analysis.

3.3.2.2 α-value optimisation for cluster core extraction

The 14 observed clusters (Figure 3.4) can generally be divided into three categories: up-regulated over the time course (clusters 2, 5, 9, 10 and 12), down-regulated with time (clusters 3 and 11) and clusters showing a distinct peak in expression at a specific time point during infection (cluster 4, 6, 7, 8, 13, and 14). Cluster 1 displays a unique expression pattern. Genes in this cluster show immediate strong down regulation up to 4 hpi, which is then released and restored up to levels of non-infected cells in 14 hours after infection. To identify a sensible value for α, the core value cut-off, cluster number 6 was filtered based on different α-values (0.75, 0.8, 0.85 and 0.9), subjected to GO term analysis and screened for TFBS enrichment in the promoter regions (-500 to 100 bp around TSSs). As mentioned above, mCMV is known to rapidly induce activation of NF-κB after infection (Benedict et al., 2004), resulting in a cluster of NF-κB responsive genes peaking in expression between 1-2 hpi (Marcinowski et al., 2012). Cluster 6 remarkably resembles the previously described expression kinetic of this NF-κB induced cluster and indeed, for all α-values, the most significant GO term associated with this cluster was “inflammatory response”. A value of 0.85 gave the lowest p-value, good enrichment and a high number of genes associated with this biological function (Figure 3.5.a). The NF-κB binding site (Figure 3.5b) could be detected as the top hit for all four values screened. The highest enrichment with significant p-values could be observed for an α-value of 0.8 (Figure 3.5.b). Since this value also gave reasonable results in the GO analysis for cluster 6, this value was consistently set as the threshold for all 14 clusters. This resulted in a minimum of 97 genes for cluster 14 and a maximum of 424 genes for cluster 12 (Figure 3.5c).

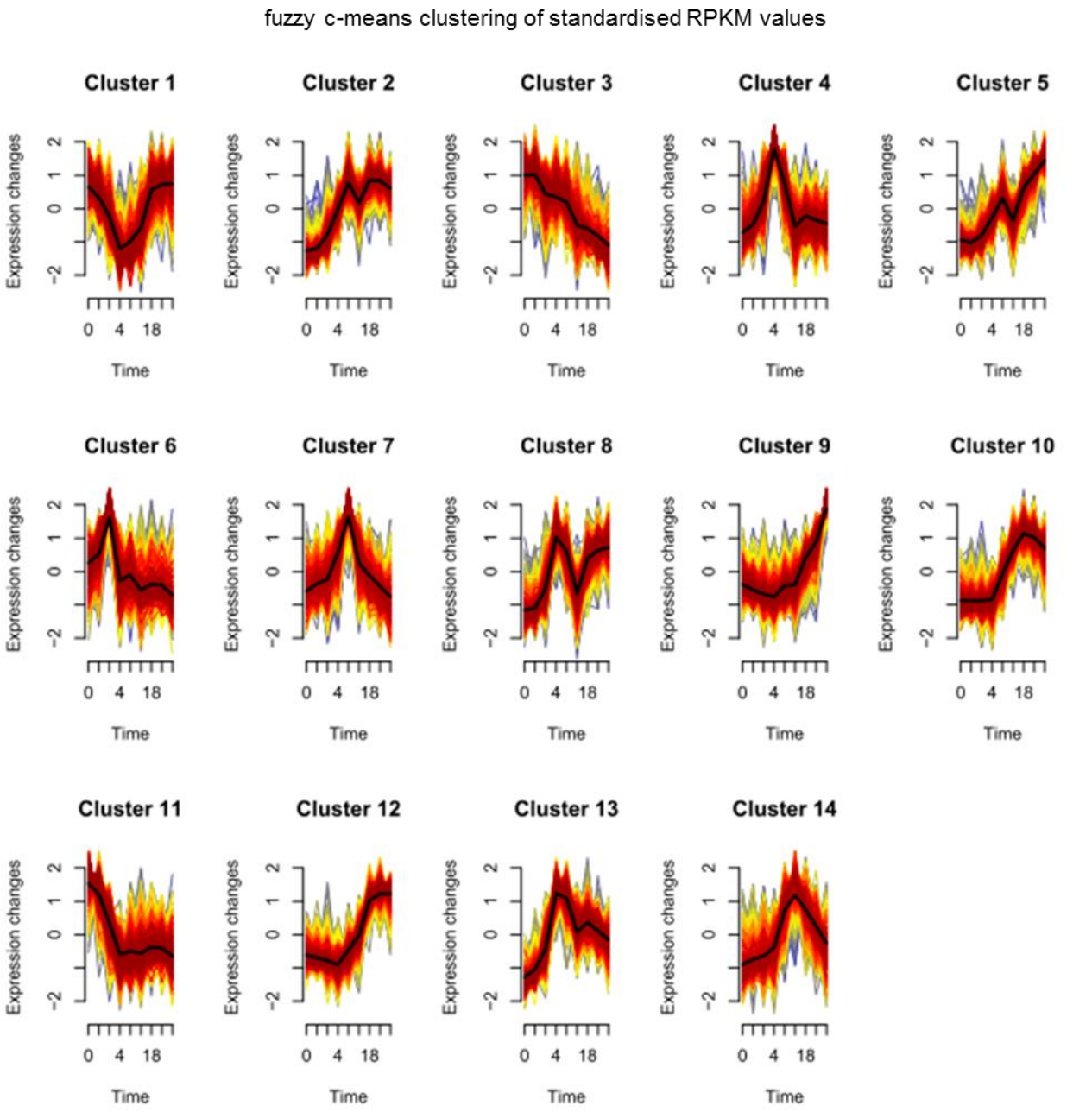


Figure 3.4 | Fuzzy c-means clustering results

Soft clustering, conducted with the Mfuzz packages, of averaged RPKM values for transcripts that possessed greater than 0.7 RPKM values in at least one of the time points in at least one of the replicates. Blue and yellow coloured lines correspond to genes with low cluster membership values; orange and red coloured lines correspond to genes with high cluster membership values. Time point 0 correspond to non-infected NIH-3T3.

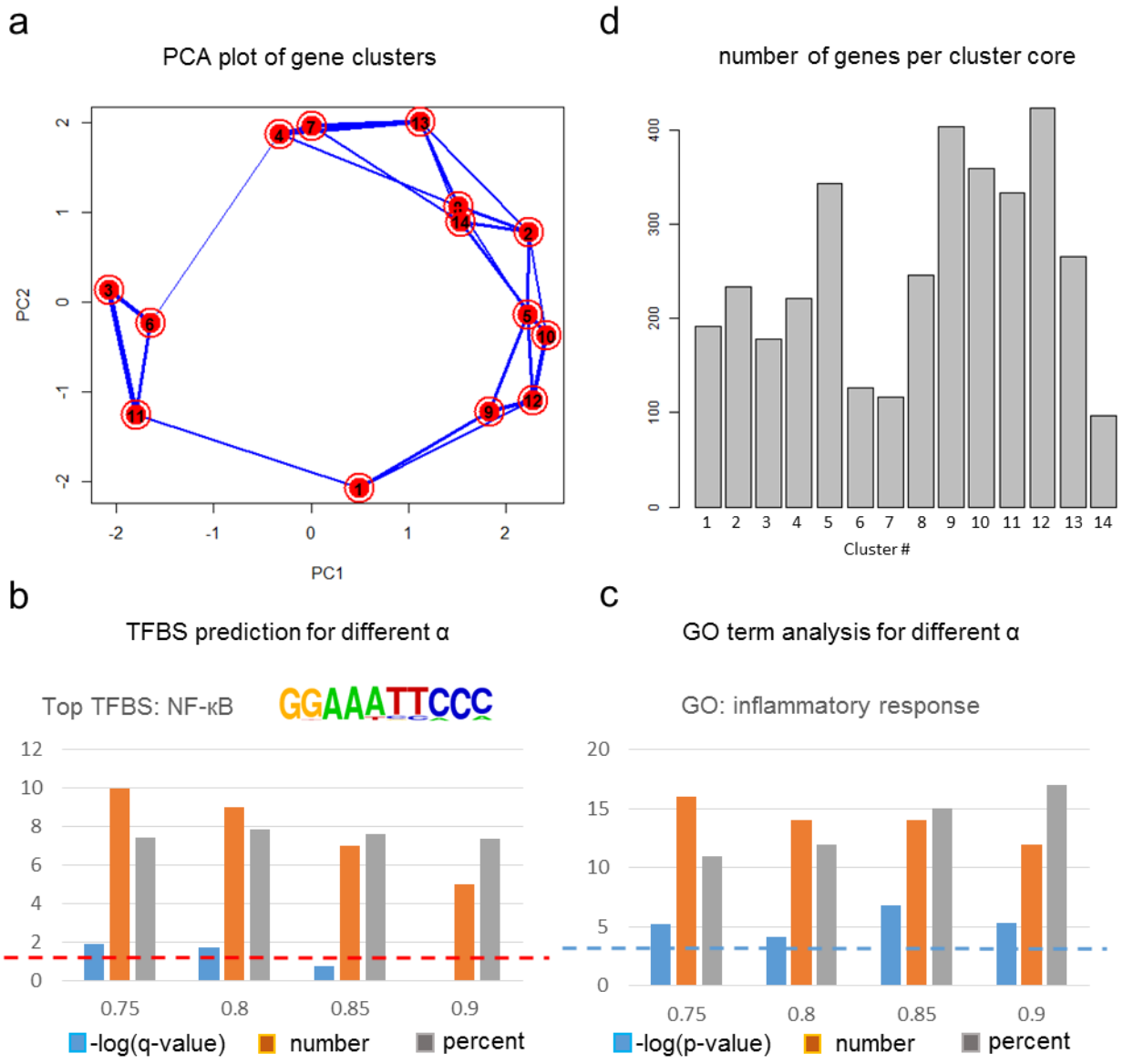


Figure 3.5 | Fuzzy c-means cluster α-core optimisation.

(a) Plot showing the first principal component plotted against the second principal component for the c=14 clusters. Different cores of the obtained cluster 6 were extracted using different values of α and subjected to **(b)** *in silico* TFBS prediction using HOMER and **(c)** GO term analysis. An α-core value of 0.8 gave the best enrichment for the TFBS prediction and significant GO terms for cluster 6 which served as training dataset. **(d)** Barchart displaying the obtained number of genes with an α-core value of 0.8. Exact number can be taken from Table 3.2.

3.3.2.3 Functional analysis of cluster cores

Quite strikingly, only one (cluster 10) out of the 14 clusters, showed no enrichment for GO terms or TFBS, demonstrating the power of the soft clustering approach. Cluster 10 contains 359 genes in its core and shows low expression up until 3-4 hpi followed by delayed up regulation until 17-18 hpi. A similar trend could be observed for cluster 12, which showed over-representation of genes involved in mitotic cell cycle progression and cell division (Figure 3.6a). Genes in cluster 12 stayed induced at the two very late time points of infection, whereas genes in cluster 10 were slightly counter regulated towards the end of the time course. The clusters

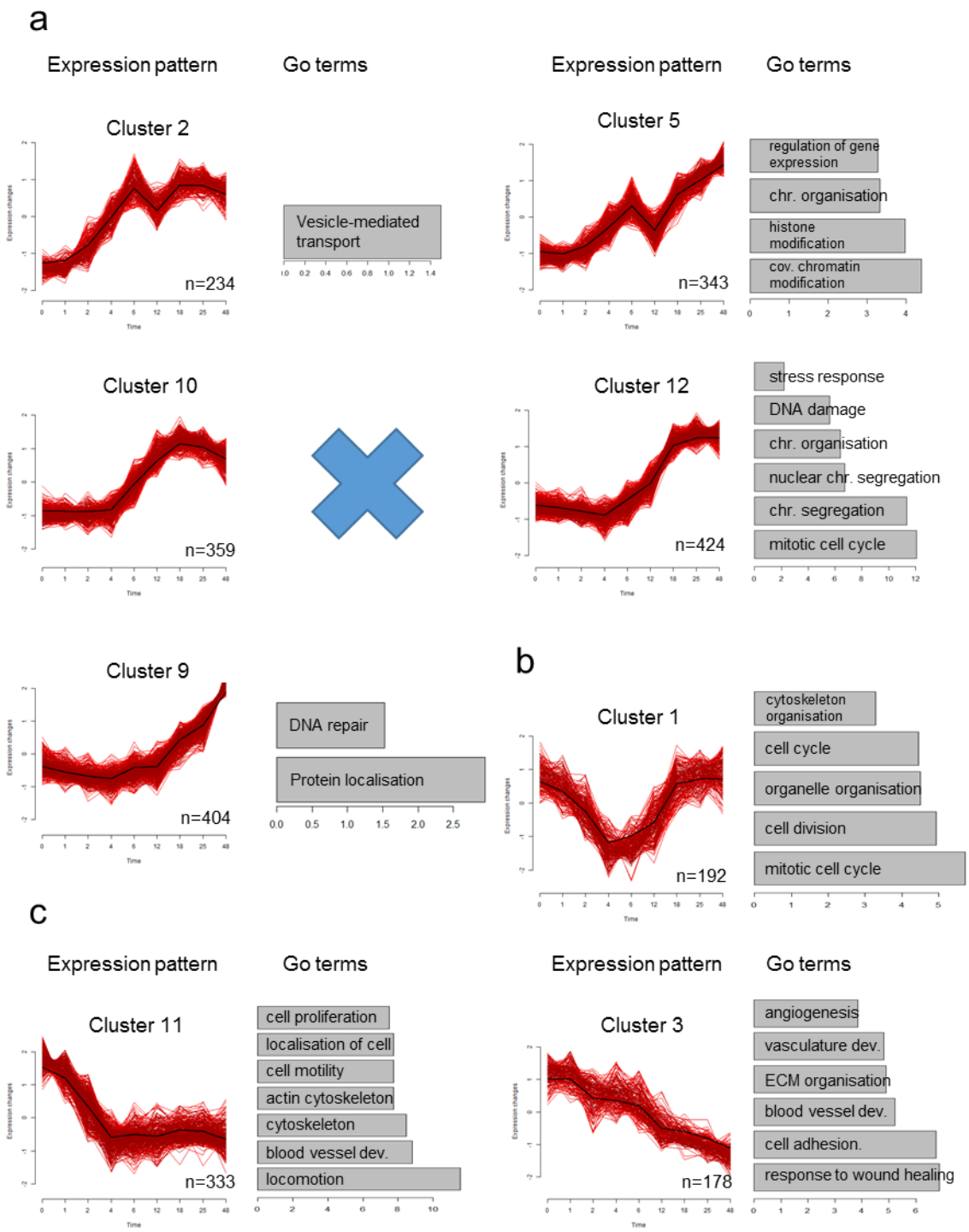
2 and 5 showed immediate and constant up-regulation early in infection, until they displayed a drop between 11-12 hpi. Genes from cluster 2 recover their expression levels between 5-6 hpi and seem to plateau at this stage, whereas genes from cluster 5 constantly increase their expression until the end of my measurements (Figure 3.6a). Genes in cluster 5 seem to be involved in mediating covalent chromatin modifications. Over-representation of binding sites for Elk1 and Elk4 could be observed in the promoter regions of genes from cluster 2 (Figure 3.6e). Genes in cluster 9 showed a delayed induction, not occurring before 11-12 hpi, but then steadily increase in expression levels and were involved in DNA repair.

In agreement with the fact that down-regulation of genes upon infection seems to be a less prominent feature of lytic mCMV infection (Figure 3.2a), I could only observe two different clusters with distinct reduction in expression. Very rapid down-regulation, until reaching a minimum at low levels of expression at 3-4 hpi, was characteristic of genes in cluster 11, while genes in cluster 3 showed slightly delayed, but then rather sustained down-regulation throughout the entire course of infection (Figure 3.6b). Rapid downregulation was characteristic for genes relating to the actin cytoskeleton, but also for genes involved in response to growth factors. Genes responsible for biological adhesion, wound healing, blood vessel development displayed delayed kinetics of transcriptional reduction. Even though, genes from cluster 1 did not strictly fit the category of downregulated genes, they displayed initial down-regulation in the first 4 hours of infection, followed by recovery back to starting levels by 17-18 hpi, where they remain until the end of the time course experiment (Figure 3.6c). Although these genes did not seem to be regulated by a specific TF, they are enriched for genes that are important for microtubule-based processes such as cell division.

Finally, the last class of cluster had in common, that genes peaked in expression levels at different time points during infection, followed by strong counter-regulation. Cluster 14 consists of genes peaking between 11-12 hpi, which are involved in processes associated with multi-cellular organisms (Figure 3.6d). Both, cluster 8 and 13, peaked between 3-4 hpi, followed by strong reduction in expression levels until 18 hpi, when they seem to recover slightly. Only genes in cluster 8 can recover the maximum expression, detected earlier in infection. Both seem to be related to RNA; cluster 8 is involved in mRNA metabolism and gene expression, whereas cluster 13 seems to be connected to non-coding RNA (ncRNA) processes (Figure 3.6b). Binding sites for NRF1 are overrepresented within PPRs of genes from both clusters. Sp1 was observed to be cluster 8 specific, while Fli1 and GABPA were specifically enriched in promoter regions from cluster 13 (Figure 3.6e). The three remaining clusters peaked at different time points early in infection, followed by strong and continuous counter-regulation. Cluster 7 featured a strong induction peak between 6-7 hpi and genes within this cluster showed over-representation of c-Myc binding sites. A prominent peak of

transcriptional activity for cluster 4 was observed between 3-4 hpi. Genes residing in this cluster were associated with ncRNA processing and binding sites for YY1 and IF can be detected more frequently than expected by random. Very early and strong induction between 1-2 hpi was characteristic for genes in cluster 6. This was followed by strong and rapid counter-regulation, back to and even below baseline levels. Genes in this cluster are significantly enriched for the association with signalling pathways in response to inflammation, but also the regulation of proliferation. In cluster 6, uniquely over-represented binding sites for NF-κB were found.

In summary, by employing 4sU-Seq throughout the entire kinetic of lytic mCMV infection, I could measure the dynamic changes in transcriptional activity of NIH-3T3 fibroblasts upon mCMV infection. Soft clustering of the data revealed multiple functional clusters, possessing distinct overrepresented TFBS, which correlate well with the assigned functions of the annotated clusters.


a
Expression pattern
Go terms
Expression pattern
Go terms
Cluster 2
n=234
Vesicle-mediated transport
Cluster 5
n=343
regulation of gene expression
chr. organisation
histone modification
cov. chromatin modification
Cluster 10
n=359
Cluster 12
n=424
stress response
DNA damage
chr. organisation
nuclear chr. segregation
chr. segregation
mitotic cell cycle
Cluster 9
n=404
DNA repair
Protein localisation
b
Cluster 1
n=192
cytoskeleton organisation
cell cycle
organelle organisation
cell division
mitotic cell cycle
c
Expression pattern
Go terms
Expression pattern
Go terms
Cluster 11
n=333
cell proliferation
localisation of cell
cell motility
actin cytoskeleton
cytoskeleton
blood vessel dev.
locomotion
Cluster 3
n=178
angiogenesis
vasculature dev.
ECM organisation
blood vessel dev.
cell adhesion.
response to wound healing

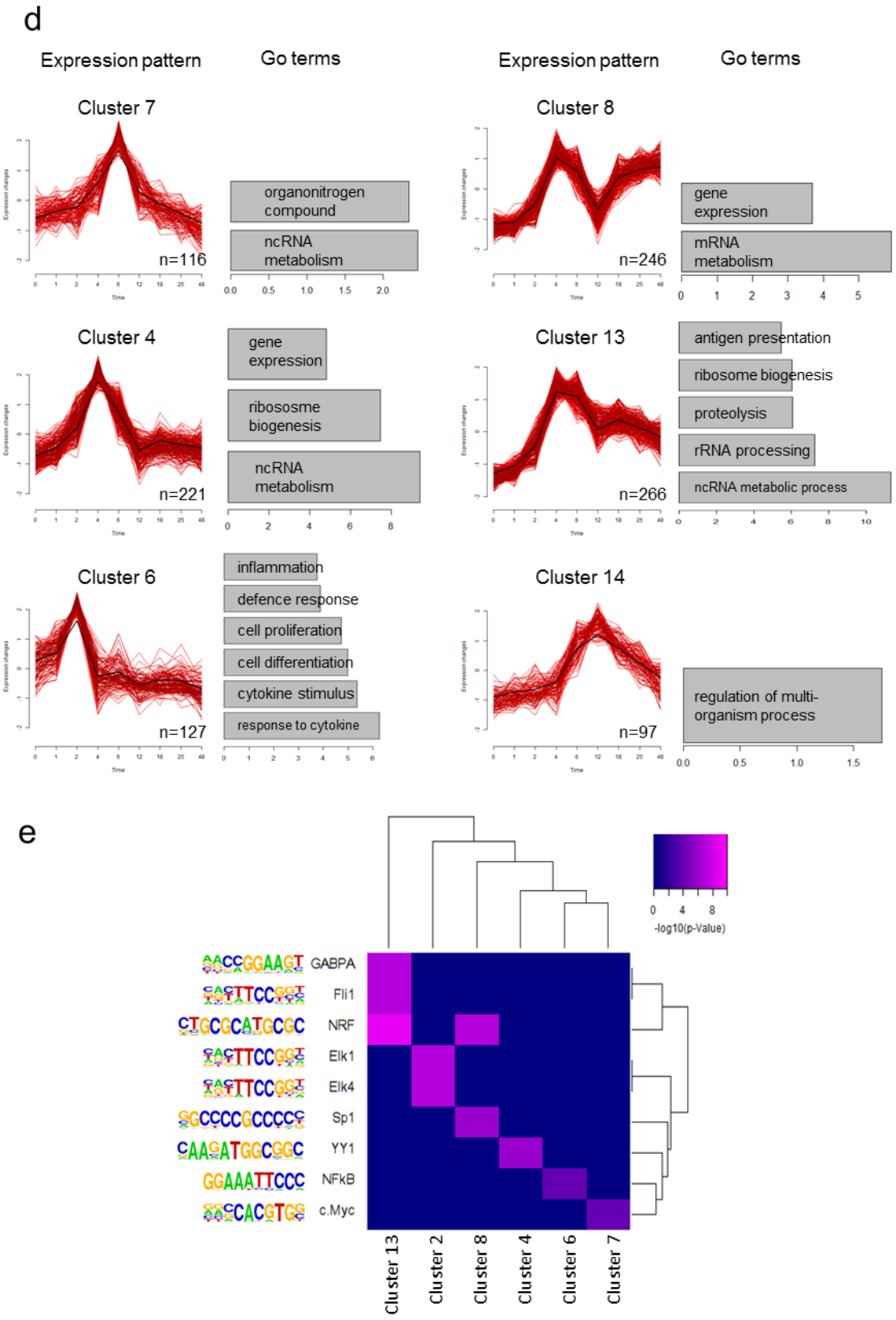


Figure 3.6 | Functional annotation of fuzzy c-means cluster cores.

Kinetic and functional GO term annotation for the obtained core clusters for **(a)** generally upregulated clusters, **(b)** genes belonging to cluster 1, **(c)** generally down-regulated clusters and **(d)** clusters that show a distinct peak in expression at a certain time point during infection. **(e)** Heatmap showing the enrichment of *in silco* TFBS prediction

using HOMER, based on the region -500 bp to +100 bp around the TSS, including the position weight matrices obtained from the JASPAR database.

3.3.2.4 Comparison with previously published clusters

A previously published study (Marcinowski et al., 2012) employed 4sU-labelling of newly transcribed RNA during the early hours (1-2 hpi, 2-3 hpi and 5-6 hpi) of lytic mCMV infection and subjected it to microarray analysis. They could identify, by manual clustering based on 2-fold changes compared to non-infected cells, 5 distinct gene cluster with characteristic differential expression, which were enriched for specific TFBS within their promoter regions (-500 bp to +100 bp around TSS). One drawback of this study was the restriction to time points early in infection. The behaviour of genes from the individual clusters past 6 hpi remained elusive.

Firstly, to see if the initially induced and then counter-regulated clusters (cluster 1 and 2) stay at low expression levels at later time points in infection and if the repression of their described clusters 4 and 5 was a constant feature of the infection, I plotted the expression levels of the previously described genes in my data. Furthermore, the expression profile of the previously described “c-Myc” cluster (cluster 3) was of interest. The TF protein itself was described as persistently expressed at least up until 12 hpi, suggesting induction of genes in the previously described cluster 3 beyond the measured time points. Beanplots of log2 RPKM value fold changes compared to non-infected cells for genes from the Marcinowski clusters obtained from my data are depicted in Figure 3.7a. I found that the general cluster-specific trends observed in the previous study was reproducible, with the exception of a couple of genes with differing temporal expression profiles between the two studies.

For example, cluster 1 and 4 show delayed kinetics in my hands, compared to the previously published results. Cluster 1 seems to peak at 3-4 hpi as opposed to 1-2 hpi, as described earlier. The inverse is observed for genes in cluster 4. Down-regulation appears to be delayed and only becomes apparent at 3-4 hpi. The delayed induction, followed by constant upregulation, of c-Myc regulated genes (cluster 4) is in agreement with the previously reported expression levels and the published levels of the TF itself. Astonishingly, a recovery of the expression levels of genes belonging to cluster 2 was observed by 17-18 hpi and was maintained in the last two time points too. A similar, but actually contrary, picture was observed for about half of the genes from the delayed but then constantly down-regulated genes (cluster 5). Expression levels of some genes of cluster 5 started to increase again by 11-12 hpi, until they reached non-infected baseline levels again by 17-18 hpi, whereas the other half of the genes showed even further reduction in expression. These observations suggest that extended sampling over time after infection is necessary to obtain a comprehensive view on transcriptional changes upon infection. This is highlighted by some of the previously published clusters, which split into sub-clusters when followed over prolonged periods. Moreover, it is

apparent in my data, that the very early time point 0-1 hpi is too early to detect any changes, not even for the most dramatically altered gene expression profiles from the five clusters described. Finally, the difference in expression kinetics of clusters 1 and 4, with very strong and instant transcriptional changes in the previously published study and delayed kinetics in my datasets might be attributed to differences in the infection protocol used. For example, slightly different viral loads could account for shifted expression kinetics.

Despite minor differences the observed expression kinetics, functional annotations and the identified TFs showed strong similarities between the two studies. This can be illustrated by examining fractions of genes within the previously published (old) clusters, which are also contained within the cores of the newly described clusters (Figure 3.7b) and *vice versa* (Figure 3.7c). The most pronounced overlap can be seen between the old cluster 4 and the new cluster 11, with 40 % of the old cluster 4 genes lying in the new cluster 11. Both clusters show immanent and sustained down regulation during the first 6 hours of infection and are enriched for genes in actin filament based processes and morphogenesis. Around 12 % of transcripts from the old cluster 4 are now belonging to cluster 3 and 6, with 6 % each. The new cluster 3 shows down regulation as well, less dramatic but rather constant throughout the complete time course. The overlap with the new cluster 6 is surprising at first. This cluster showed strong induction as early as 1-2 hpi, followed by very strong suppression throughout the rest of the measurements, whereas immediate and strong suppression was characteristic for the new cluster 11. The common characteristic, however, is the strong suppression post 2 hpi. One has to keep in mind that the published data were obtained by Microarray analysis and clusters were defined based on 2-fold changes of non-standardised values. Thus, comparison between different transcripts is difficult and will further lead to the difference in sensitivity at high expression levels (Draghici et al., 2006). By plotting the averaged RPKM values for these genes (Figure 3.7c), it becomes obvious that they indeed are being more expressed 0-1 hpi compared to non-infected samples. The second biggest overlap with 23 % can be observed for the old cluster 1 and the new cluster 4. Genes within those clusters show a prominent peak of expression very early upon infection, and are associated with the GO term of “regulation of gene expression”. A further 12 % of the old cluster 1 are now part of the new cluster 6, which displays the exact same kinetics, a pronounced peak of transcription 1-2 hpi, is enriched for genes involved in the inflammatory response and possesses more NF-κB binding sites than expected by chance. One would have expected to observe the opposite picture of enrichment for genes from the old cluster 1, with a greater overlap with the new cluster 6, due to the similarity in expression kinetics. I can only speculate that this again might reflect differences in the infection protocol, resulting in different viral loads between the two studies. As mentioned earlier, there is a striking similarity between the old cluster 5 and the new cluster 1. Both are being down regulated early upon mCMV infection until 4-6 hpi. Gene expression levels recover

subsequently and reach levels comparable to ones in non-infected cells at 17-18 hpi. Hence, the observed overlap with 15 % is the third largest overall. The overlap of the old cluster 2 with both, the new cluster 4 and 8, highlights the importance of extending the time course. All three clusters show a peak of expression between 3-4 hpi, followed by counter regulation 5-6 hpi. Genes from the new cluster 4 remain repressed, whereas genes from the new cluster 8 recover their expression levels after reaching the minimum at 11-12 hpi. These clusters seem to be associated with non-coding or coding RNA metabolism, respectively. The old cluster 3 is less well represented in the new data. The biggest overlap, with 8 % each, could be observed with the new clusters 5 and 10. Both are quite large clusters with delayed induction at around 5-6 hpi, which is constant for the new cluster 10, whereas the new cluster 5 displays a dip in expression between 11-12 hpi. The new clusters 2, 5, 7 and 8 are only weakly represented by the old clustering approach, while cluster 9, 12, 13 and 14 have not been described before.

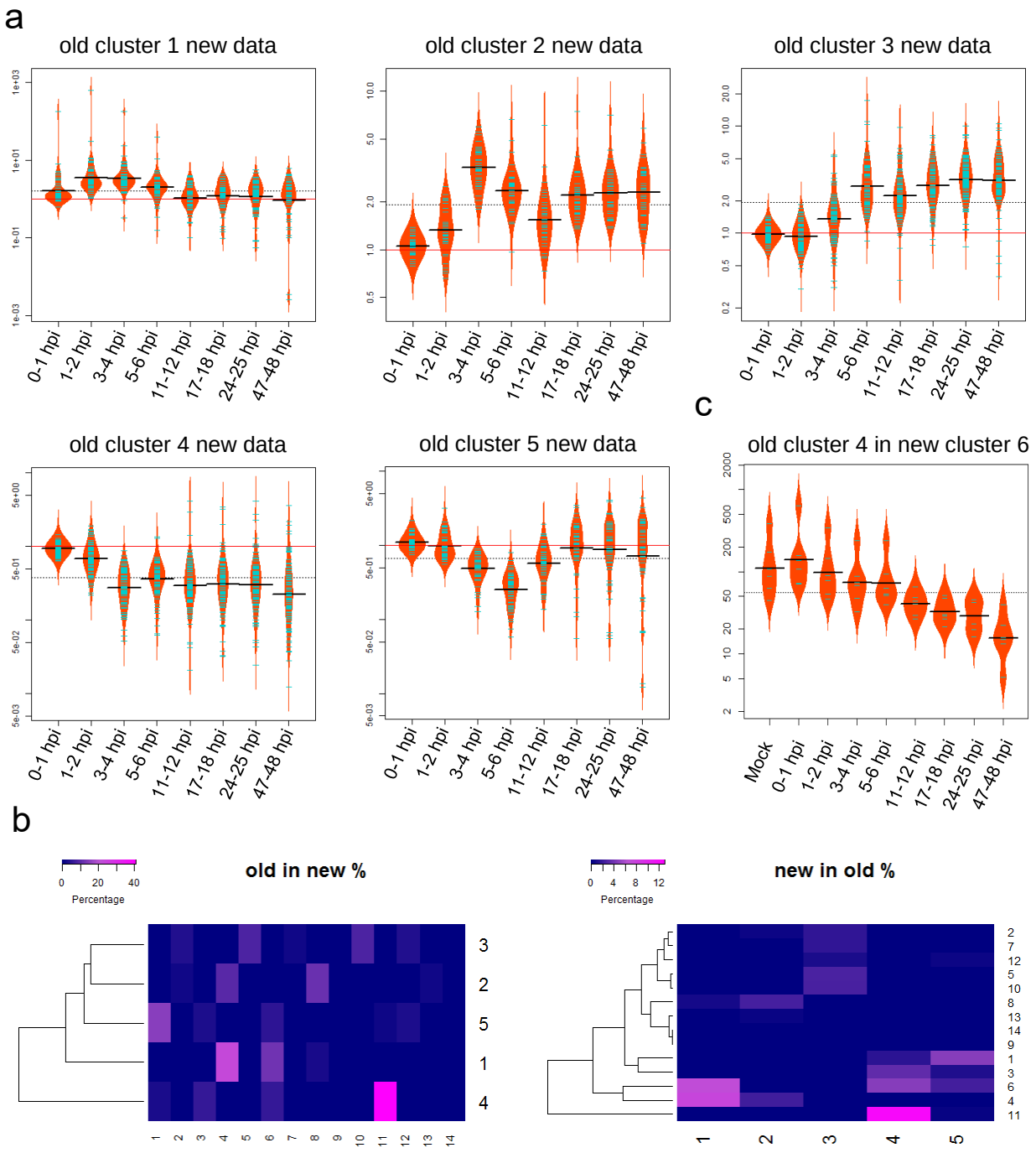


Figure 3.7 | Comparison of clusters obtained by soft clustering with published clusters

(a) Beanplots showing averaged log2 expression data fold changes in RPKM compared to non-infected cells from both of my replicates from the previously described clusters (Marcinowski et al., 2012). Green lines are depicting a one-dimensional scatterplot of these log2 fold changes, which is converted into a density distribution, mirror-imaged and displayed in orange. Solid black lines are indicative of the mean of the distribution, whereas red lines are representing the mock-infected levels. The overall mean between all time points is depicted as a dashed black line. **(b)** Heatmaps indicating the percentage of overlap between the newly defined and the previously described clusters (left) and *vice versa* (right). **(c)** Bean plots depicting the log2 expression levels detected in my experiments for genes that were previously described to be instantly down regulated but now appear to peak in expression 1-2 hpi prior to down-regulation.

In summary, fuzzy c-means is a powerful tool to analyse the dynamic changes of newly transcribed RNA upon lytic mCMV infection and enabled the identification of 13 clusters significantly associated with a cellular function. Furthermore, by extending the time course to nine time points, I provided a more comprehensive picture and could describe so far unknown kinetics of host gene expression.

3.3.3 Genome-wide chromatin accessibility measured by ATAC-Seq

3.3.3.1 Establishing ATAC-Seq on MCMV infected cells

Chromatin accessibility is indicative of the status of regulatory sequences (Shlyueva et al., 2014; Thurman et al., 2012) and nucleosome positioning is important at many genomic elements, particularly at promoters (Jiang et al., 2009). In order to assess the accessibility of promoters from the different clusters and to define proximal regulatory regions (PRRs), rather than using the -500bp to 100bp definition of core promoters, I generated ATAC-Seq libraries from three different biological replicates matching the end time points of the 4sU labelling (1, 2, 4, 6, 12, 18, 25 and 48 hpi as well as non-infected NIH-3T3). Library concentrations and size distributions were checked by KAPA qPCR and Bioanalyzer analysis. All libraries displayed the expected periodicity of ~200 bp (Figure 3.8), which is caused by preferential insertion of the sequencing adapters by the transposase between nucleosomes.

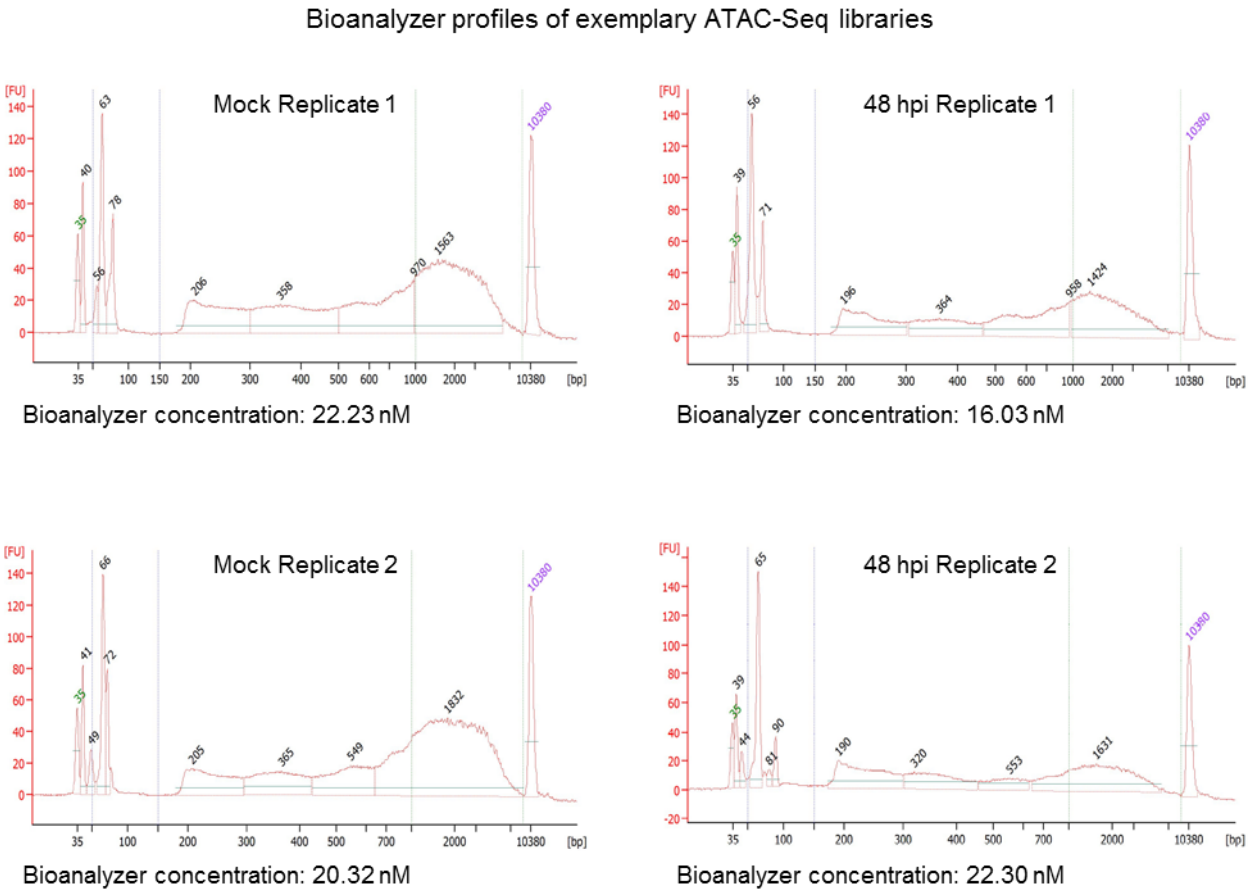


Figure 3.8 | ATAC-Seq library size distribution and concentrations pre-sequencing

Bioanalyzer profiles for exemplary ATAC-Seq libraries following final amplification and purification, showing clear periodicity at ~200 bp, for two of the three replicates in non-infected NIH-3T3 and 48 hpi, MOI = 10. Library concentrations are reported from the Bioanalyzer profiles for the entire range >150 bp.

Cambridge Genomic Services sequenced the libraries on a NextSeq500 platform, using the 75 bp paired-end mid-output kit, on one lane per replicate. Sequencing reads were processed and mapped to the mm10 mouse genome built, including the viral genome as an extra chromosome (as described in 2.14.1). The obtained, mapped and valid read-pair numbers can be taken from Table 3.2 below.

Table 3.2 | ATAC-Seq mapped read-pair numbers

Libraries were sequenced with 75 bp paired end sequencing on the NextSeq500 platform, using one lane per replicate.

Time point (TP)	Replicate 1	Replicate 2	Replicate 3	total per TP
Mock	16,685,931	13,721,167	11,055,353	41,462,451
1 hpi	14,111,151	12,091,324	6,943,008	33,145,483
2 hpi	18,984,686	8,377,607	8,515,856	35,878,149
4 hpi	17,413,579	11,453,151	14,842,512	43,709,242
6 hpi	16,097,026	14,399,746	12,335,193	42,831,965
12 hpi	14,611,438	14,977,317	27,423,948	57,012,703
18 hpi	16,531,502	17,741,513	21,853,285	56,126,300
25 hpi	12,356,198	12,208,788	24,483,994	49,048,980
48 hpi	17,016,777	9,769,490	10,795,471	37,581,738
total per Rep	143,808,288	114,740,103	138,248,620	396,797,011

3.3.3.2 Genome-wide ATAC-Seq chromatin accessibility

To identify open chromatin regions from ATAC-Seq data, I called peaks using MACS2 (as described under 2.14.2), on individual time points and on time points pooled across replicates. Overall, good agreement between replicates could be observed, in non-infected cells (Figure 3.9a) but also at late stages of infection (Figure 3.9b), with more than 75 % overlap of all peaks between replicates. Even at different stages of infection, the overlap between detected peaks was quite high (Figure 3.9c), although a substantial amount of time point specific peaks could be observed, with the non-infected samples having the highest unique and highest absolute number in peaks. Differential peak calling in ChIP-Seq and ATAC-Seq type data is a current problem in medical and biological research and has not been extensively addressed yet. There are several tools for comparing two biological conditions (Liang & Keles, 2012), but a sophisticated way of analysing ATAC-Seq time course data is missing to date.

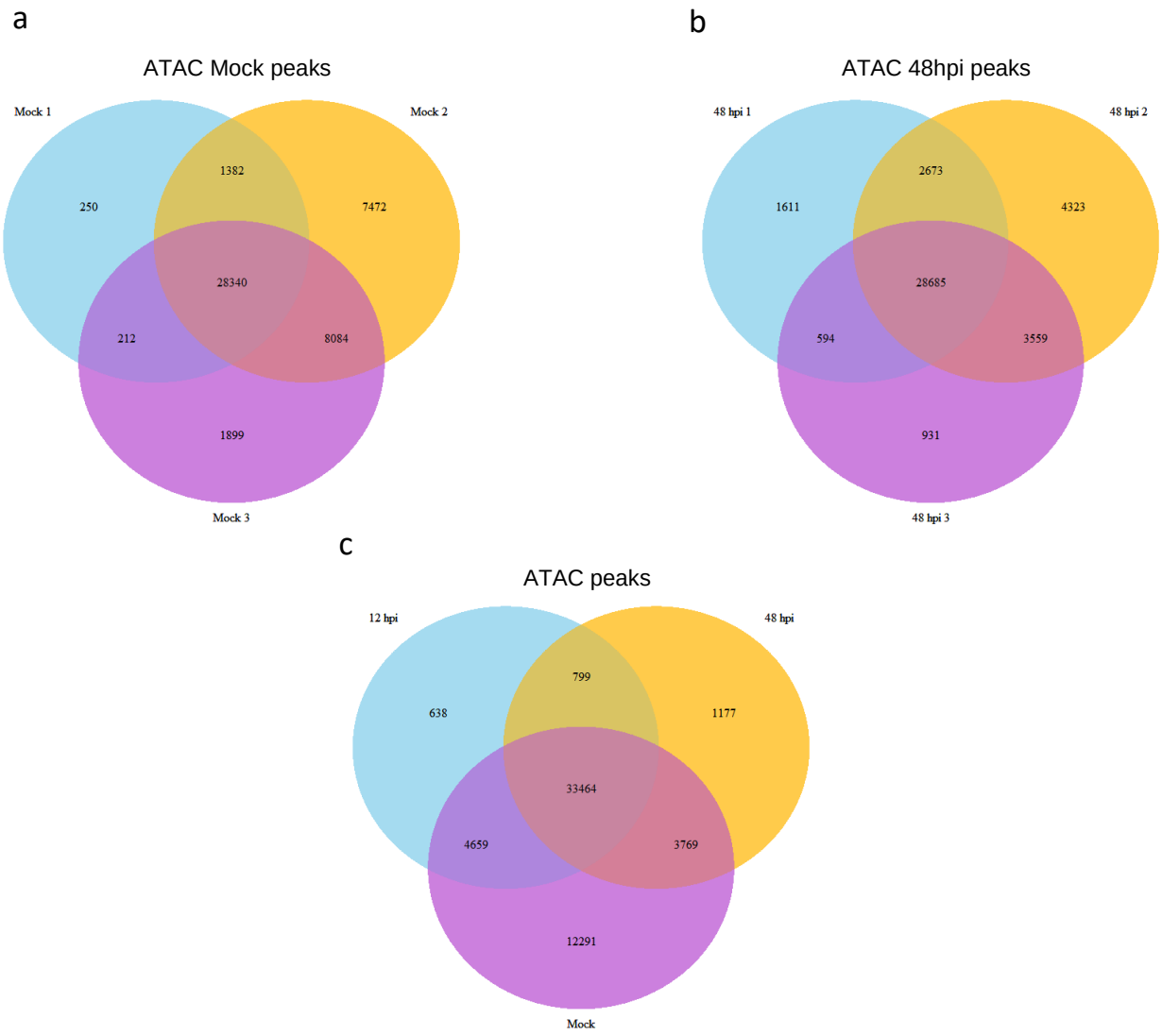


Figure 3.9 | ATAC-Seq peak calling

ATAC-Seq peaks were called using MACS2 and the overlap of peaks between the three replicates is depicted for **(a)** non-infected cells and **(b)** cells 48 hpi. **(c)** The union of all time points specific peaks was formed and the overlap between exemplary time points of infection is displayed. Of note, circles are not to scale.

To obtain the largest possible set of peaks, I formed the union of all peaks obtained from calling peaks on combined replicates per time point. This resulted in 48,670 peaks, which was reduced to 48,417 by removing peaks overlapping known blacklisted regions (obtained from www.encodeproject.org). The three biological replicates showed strong correlation in the number of reads at these peaks, as shown for the mock infected replicates and the 48 hpi replicates in Figure 3.10a and Figure 3.10b, respectively, with R values between 0.92 and 0.94 (Figure 3.10c). Thus, I decided to pool read-pairs for the individual time points across replicates, to increase the signal to noise ratio per time point. The finding, that there are not many major differences occurring with the ongoing infections, in terms of peak numbers detected, is supported by visual inspection of the data. The region plotted in Figure 3.1a is matching the one depicted in Figure 3.10d, now showing the absolute ATAC-Seq signal of

pooled replicates for all time points collected. Clear pileup of reads can be observed over TSSs and some intergenic regions and only regions with high coverage compared to the background are annotated as peaks.

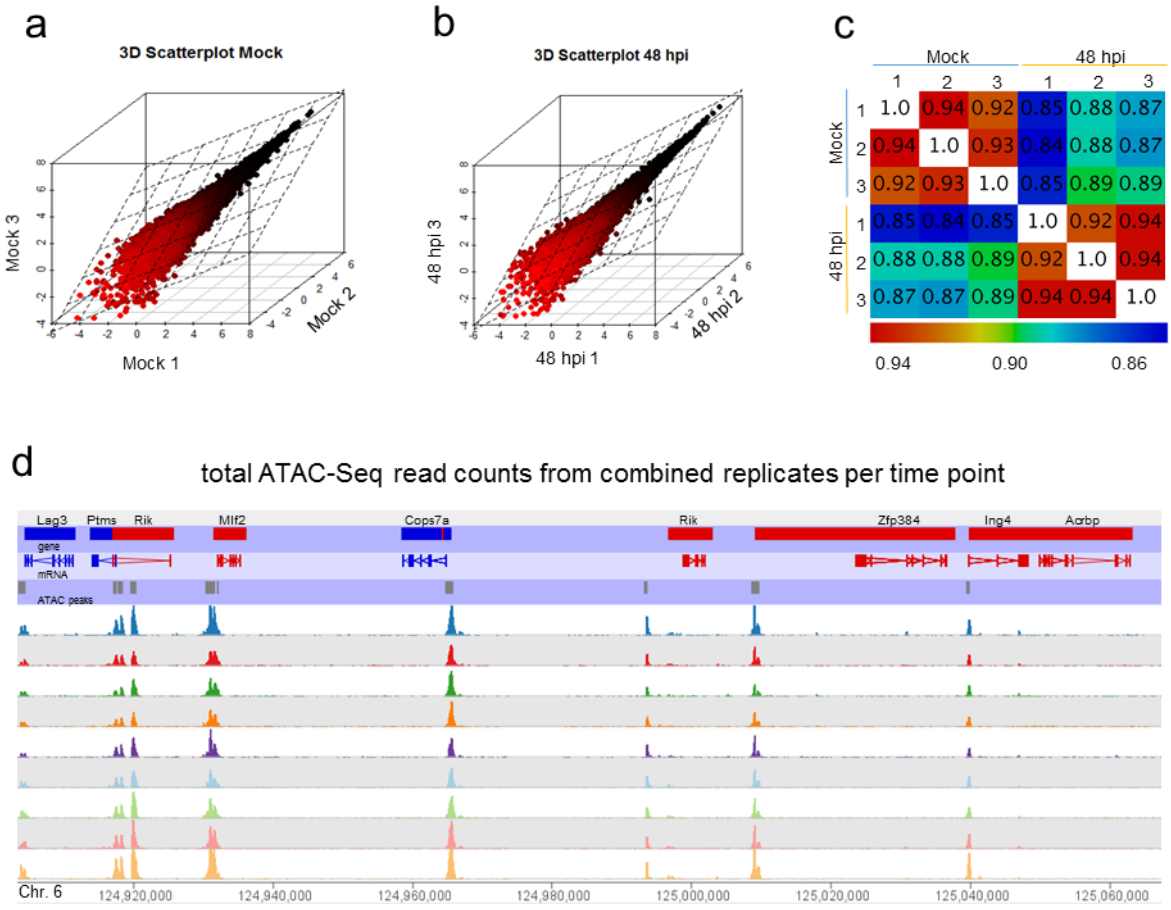


Figure 3.10 | Assessing chromatin accessibility using ATAC-Seq

Three-dimensional scatterplots depicting the correlation between the three ATAC-Seq replicates for **(a)** mock-infected and **(b)** cells 48 hpi. **(c)** Pair-wise Pearson`s correlation coefficient between all libraries obtained from mock-infected cells and cells 48 hpi. **(d)** Genome browser shot of a representative locus in the mouse genome showing total reads from combined replicates for all time points.

3.3.3.3 Accessibility surrounding TSS

Active promoters are frequently characterised by accessible nucleosome-free regions at their TSSs (Jiang et al., 2009). In order to analyse how ATAC-Seq signal at promoters correlates with gene transcription, I plotted the distribution of ATAC-Seq reads from +1.5 kb to -1. kb around the TSS of all genes for all time points (Figure 3.11a). The average of all genes, displays the most accessible region directly on the TSS. Further, immediately downstream of the TSSs a dip in accessibility, corresponding to the relatively strong positioning of +1 nucleosomes, is apparent. The most distinct ATAC-Seq signal across all genes in all time points in my data was measured within the region +1 kb to -700 bp around the TSSs, with clear local minima at the borders (Figure 3.11a). Furthermore, I separated genes into expressed and

non-expressed genes based on their expression levels in non-infected cells and further divided the expressed ones into quartiles. Examining the 1.7 kb region around their TSSs, which displayed the highest accessibility across all genes and time points, revealed an overall trend for genes with higher expression levels to have a higher ATAC-Seq signal around their TSS (Figure 3.11c). Highly expressed genes are also more likely to have a MACS2 defined peak at the 1.7 kb region around their TSS (Figure 3.11d). Even though, the genome-wide quantitative correlation for all expressed genes between transcriptional activity and accessibility around the promoter is significant and positive, it is only very weak (Figure 3.11e). Instead, I found a bimodal distribution of ATAC-Seq values around promoters (Figure 3.11f). Silent and very lowly expressed genes occupy the lower distribution of values, whilst expressed genes, past a certain threshold, are more likely to occupy the higher distribution of values. Together these data suggest that a certain level of accessibility around a genes promoter is required for expression of this gene, and this likely corresponds to the presence of a nucleosome free-region. However, ATAC-Seq signal does not correlate strongly with the level of transcription, past this threshold.

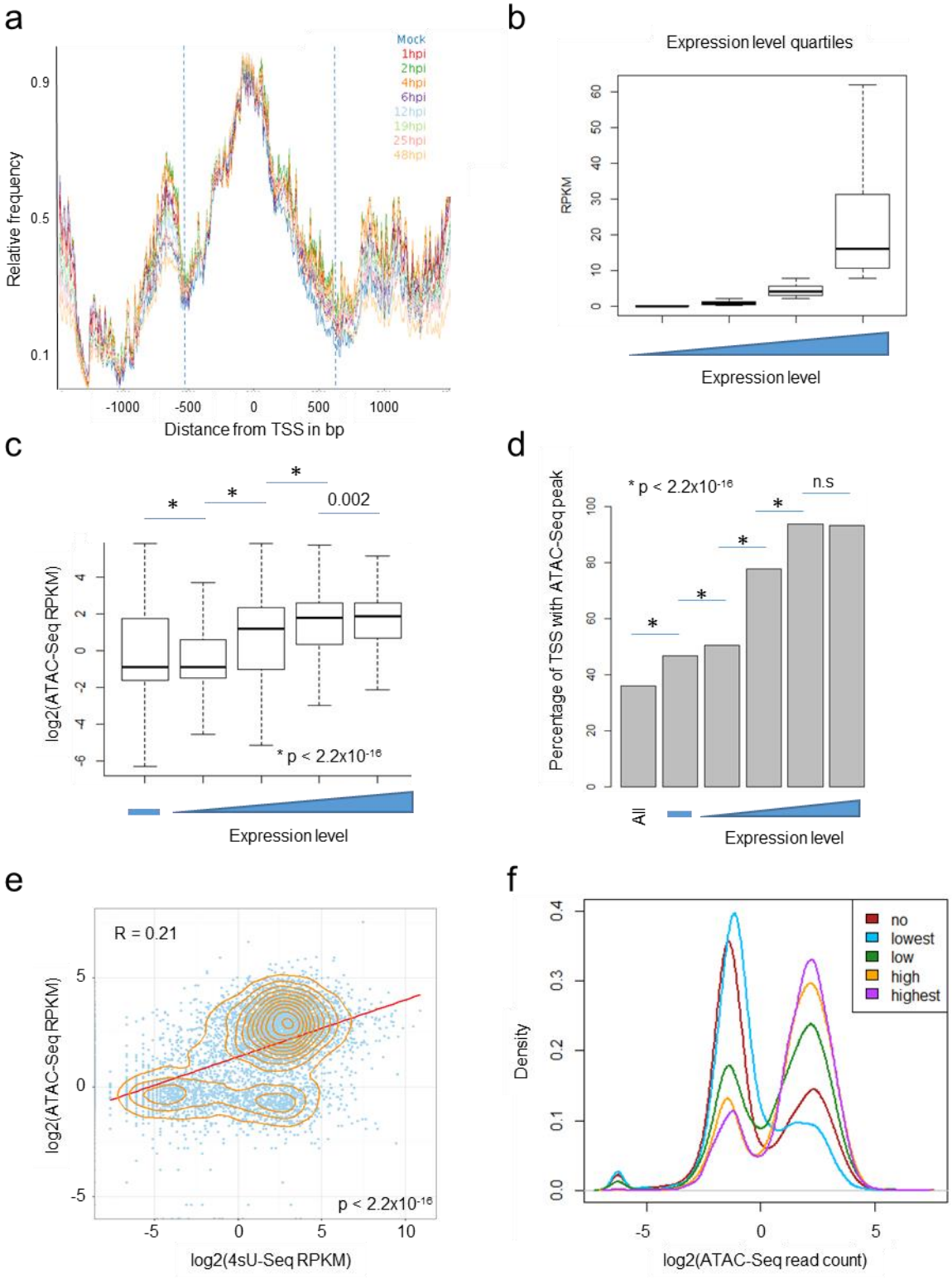


Figure 3.11 | ATAC-Seq signal around TSS

(a) Normalised cumulative distribution of ATAC-Seq read counts for all combined replicates per time point surrounding TSS. Blue dotted lines indicate the 1.7 kb used for subsequent analysis. **(b)** Genes were categorised into non-expressed and expressed genes based on an expression cut-off of 0.7 RPKM values detected in non-infected cells. Expressed genes were further subdivided into expression quartiles and the expression values for each quartile is depicted in boxplots. Boxplots show the median, interquartile range and range (outliers, defined by >1.5 times the interquartile range away from the box, are not plotted). **(c)** Total ATAC-Seq read counts

surrounding TSS (1.7 kb windows around TSS) of non-expressed genes (depicted in the first column and represented by the large blue minus) and genes in different expression quartile categories (depicted from 2nd to 5th column with increasing expression levels). Boxplots as before. P values were calculated using then Wilcoxon rank-sum test. **(d)** Percentage of TSS in each expression category (represented as above), which possess an ATAC-Seq peak called by MACS2 within the 1.7 kb window surrounding the promoter. P values were calculated using the same test as above. **(e)** Correlation between ATAC-Seq read counts within the 1.7 kb promoter regions and gene expression values (RPKM). R values corresponds to Pearson`s correlation coefficient. **(f)** Density plot showing the distribution of log2(ATAC-Seq read counts for TSSs) (1.7 kb windows) in different expression categories.

3.3.3.4 PPR determination using ATAC-Seq

In order to act on transcription, most TFs have to bind to accessible stretches of DNA. Hence, ATAC-Seq signal can be used to improve PPR of genes. So far, I have used the +500 bp to -100 bp region of genes to computationally predict TFBSs. I hypothesised, that by looking at ATAC-Seq peaks overlapping the 1.7 kb PPR region around TSSs defined earlier, a more detailed picture of where and which TFs bind will emerge. Therefore, I extracted ATAC-Seq peaks overlapping those regions and repeated the *in silico* TFBS prediction using HOMER. First, it becomes apparent, that I identified overrepresentation of specific TFBS for all clusters now, many of which are present and significant for multiple clusters (Figure 3.12a). Especially the ETS TF family (including Elk1, Elk4, Elf1, GABPA and Fli1) is highly over-represented in the PPRs of genes belonging to the clusters 2, 4, 5, 7 and 13, all of which show an induction early in infection until 6 hpi, when they start to decrease in expression until 12 hpi. The two down-regulated clusters, cluster 3 and 11, but also cluster 8, show strong enrichment for binding sites for Fos family TFs and AP-1 related TFs. I was further able to detect cluster-specific TFBS, which were not detectable when looking at larger regions of fixed length around promoters. To properly assess the differences in TFBS prediction, using either the +500 bp to -100 bp definition or the regions determined to be accessible by ATAC-Seq, I plotted the difference in p-values for the different TFs in the respective clusters (Figure 3.12b). As expected, many new over-represented binding sites could be observed (depicted in red), but also more than a handful of previously enriched binding sites were not significantly over-represented anymore (visualised in blue). Amongst these are the NF-κB binding site previously observed in cluster 6, the YY1 and IF binding sites in cluster 4 and the c-Myc binding sites in cluster 7. Nevertheless, other clusters now contain these binding sites more frequently than expected by chance. For example, the over-representation of binding sites for YY1 in cluster 4 is not significant anymore, but a significant enrichment of this motif can now be observed for the clusters 5, 9, 10 and 13, all of which show early induction and never fall back down below the expression levels of non-infected cells.

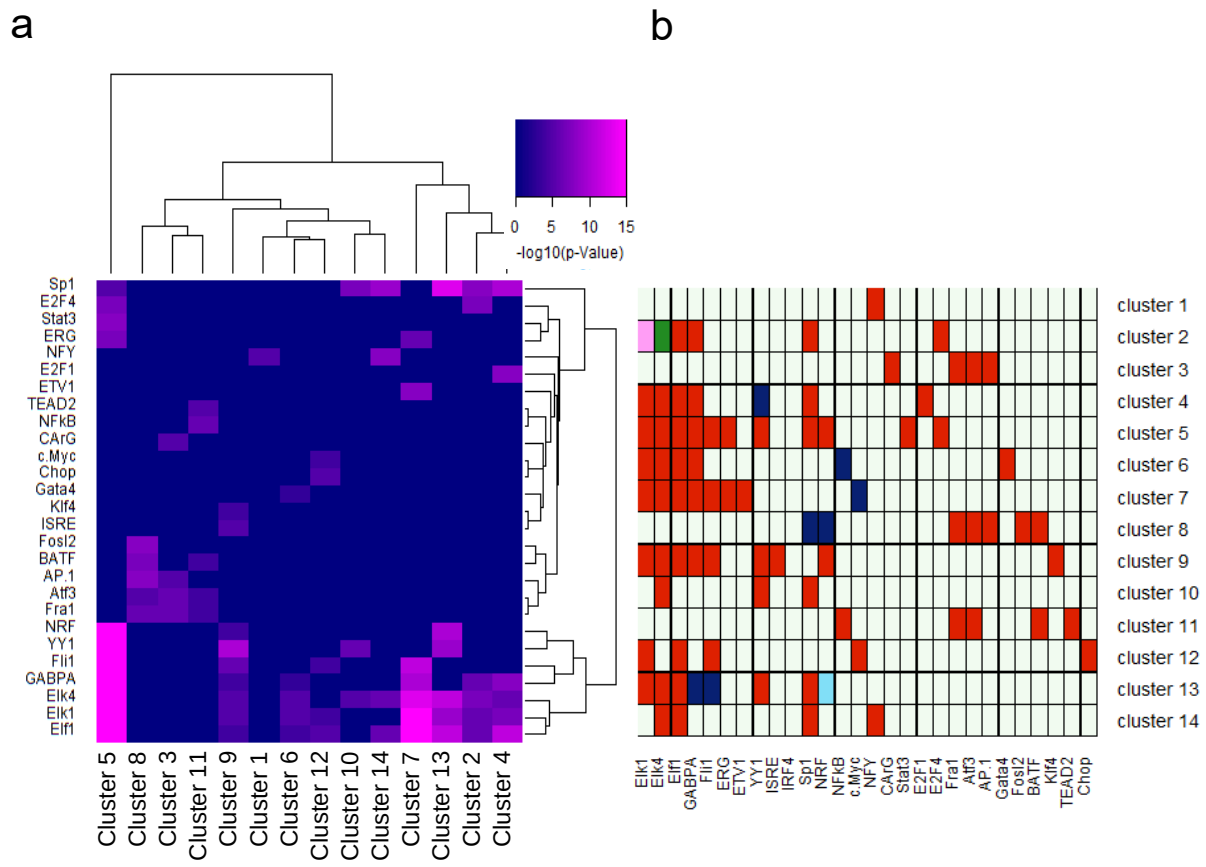


Figure 3.12 | PPRs of different gene clusters are enriched for specific TFBS

(a) Heatmap showing the enrichment of *in silco* TFBS prediction using HOMER for the given clusters, based on PPRs determined by ATAC-Seq peaks surrounding the promoters. **(b)** Heatmap indicating differences in *in silico* TFBS prediction when either using the -500 bp to +100 bp promoter definition or PPRs determined by ATAC-Seq. Detection of new motifs in red; loss of motif in dark blue; green colouring depicts no changes; decreasing and increasing p values are depicted in magenta and light blue, respectively.

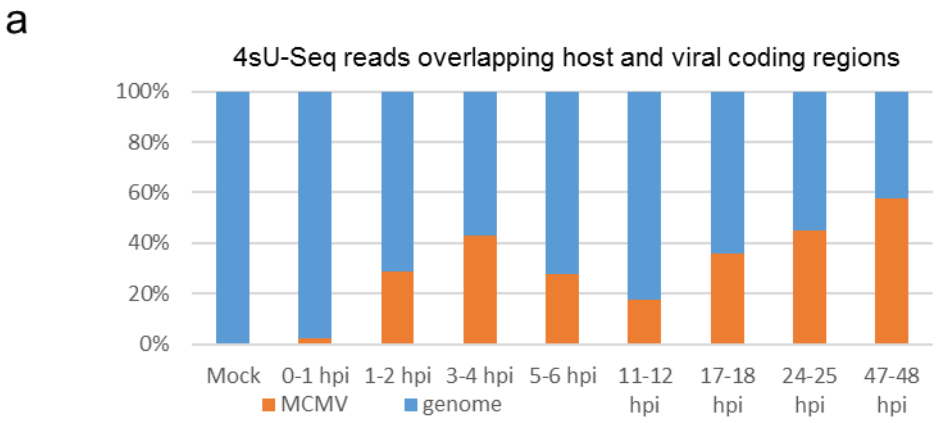
Taken together, I found that accessibility at TSS is highly correlated with transcription of the gene in a binary manner. However, no linear, quantitative correlation between transcriptional output and accessibility could be detected past a certain threshold. Nonetheless, information about nucleosome-free regions can help to improve the annotation of PPRs and therefore, increase the sensitivity of *in silico* TFBS prediction.

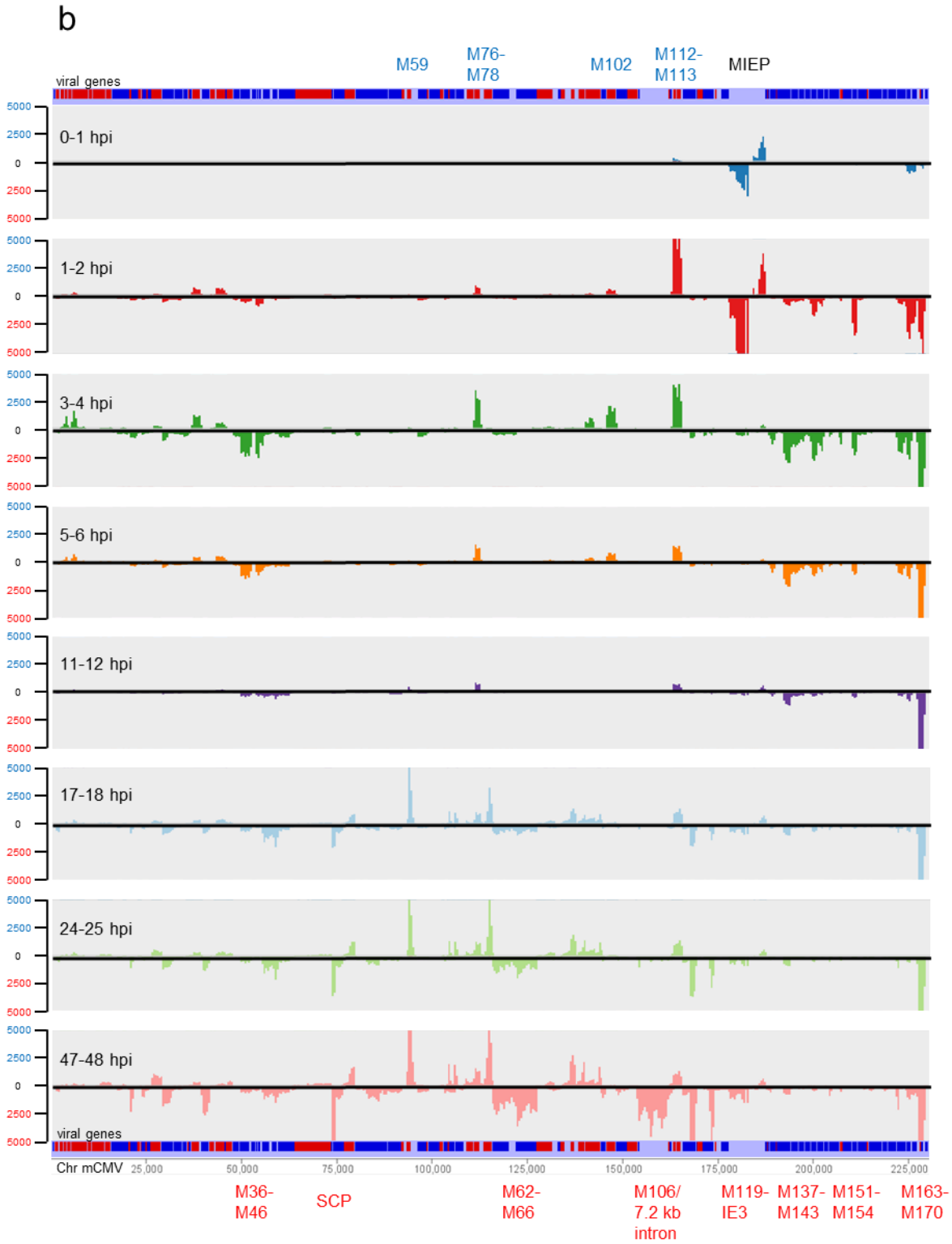
3.3.4 Viral gene expression and accessibility data

Metabolic labelling of newly transcribed RNA enables assessing changes not only in host but as well in viral gene expression, without being hindered by the large amount of virion-associated RNAs with longer half-lives. When considering only coding regions, viral transcripts accounted for ~60 % of all transcripts measured by 4sU-RNA-seq at 47-48 hpi (Figure 3.13a). Notably, the extent of viral transcripts at 1-2 hpi accounted already for 30 % of all transcripts, even increasing to more than 40 % at 3-4 hpi. Viral expression levels then steadily drop until reaching a minimum at 11-12 hpi. Just after the onset of viral DNA replication

(~15 hpi) (Marcinowski et al., 2012), viral gene expression increases again. A more detailed view of viral gene expression kinetics is displayed in Figure 3.13b. This reveals that viral gene expression is initiated as early as within the first hour after viral entry at the MIEP in a bidirectional fashion. At 1-2 hpi transcription is arising from multiple additional loci across the whole viral genome and is reaching a temporal maximum at 3-4 hpi. By 5-6 hpi, the transcription rates of many but not all viral genes are substantially reduced and reach a minimum at 11-12 hpi. Only a region over M163-M170 (3`-end of viral genome) maintains very high expression levels throughout the entire infection. The observed increase in viral gene expression observed at 17-18 hpi, just past the onset of viral DNA replication (Figure 3.13a), can be assigned to the expression of viral late genes such as M129, M59, M94 and M106, including the 7.2 kb viral intron of the viral gene M106.

Genome-wide accessibility of the viral genome, determined by ATAC-Seq, corroborated the above findings, and provided insights into chromatin compaction of the viral genome. In general, very early in infection the entire viral genome is accessible, without the display of specific peaks or less accessible regions. Interestingly, a substantial drop in accessibility was observed at 4 hpi, with accessibility of the entire viral genome continuing to drop until 25 hpi. Subsequently, the entire viral genome becomes slightly more accessible again until it reaches its late stage peak at 48 hpi (Figure 4.13c). A striking observation was the very prominent -1 and +1 positioned nucleosome at all viral promoters at 1 hpi indicated by low ATAC-Seq signals. This position effect seems to vanish as early as 2 hpi and seems to shift towards a depletion of the ATAC-Seq signal were the +2 positioned nucleosome would be expected. At later stages of infection, it appears that the +1 nucleosome is depleted genome-wide at viral promoters as ATAC-Seq signals are reinforced at the corresponding position (Figure 4.13d).




b
M59
M76-M78
M102
M112-M113
MIEP
viral genes
0-1 hpi
1-2 hpi
3-4 hpi
5-6 hpi
11-12 hpi
17-18 hpi
24-25 hpi
47-48 hpi
viral genes
Chr mCMV
25,000
50,000
75,000
100,000
125,000
150,000
175,000
200,000
225,000
M36-M46
SCP
M62-M66
M106/7.2 kb intron
M119-IE3
M137-M143
M151-M154
M163-M170

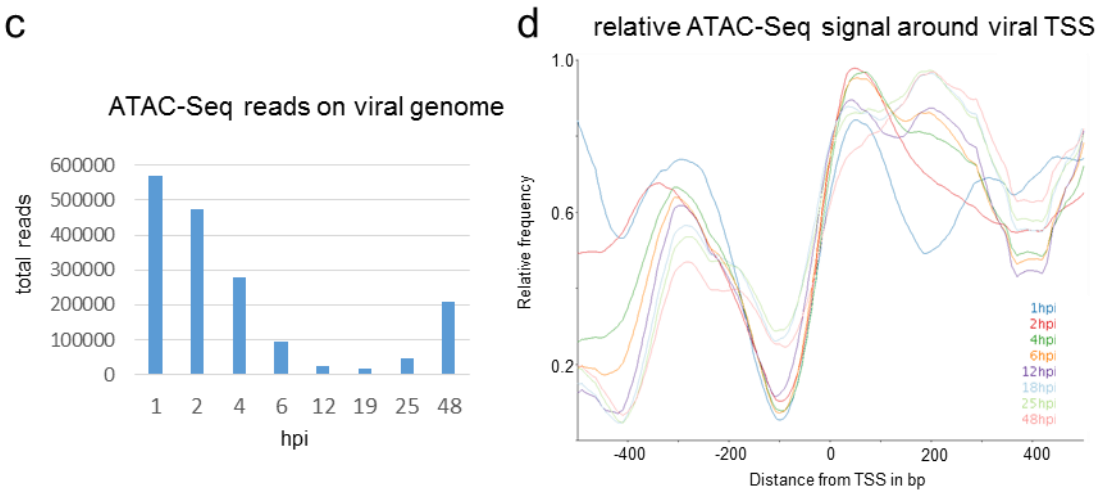


Figure 3.13 | Accessibility and transcription of the viral genome determined by ATAC-Seq and 4sU-Seq
(a) Barplot showing the relative amount of viral compared to host RNA-Seq reads falling onto transcripts for combined replicates. Absolute numbers can be obtained from Table 3.1**. (b)** Genome browser view of the viral genome at different stages of infection depicting forward (up) and reverse (down) RNA-Seq reads at 500 bp resolution**.** Gene annotations of known coding viral regions were extracted from (Rawlinson et al., 1996) and are indicated in blue for transcripts overlapping the sense direction and in red for transcripts on the anti-sense strand. The bi-directional MIEP is depicted in black. **(c)** Barplot displaying absolute ATAC-Seq read counts mapping to the viral genome at the indicated stages of infection (combined replicates per time point) **(d)** Normalised cumulative distribution of ATAC-Seq read values around viral TSS for all time points of infection (combined replicates).

In summary, applying 4sU-Seq and ATAC-Seq to viral infections furthers our understanding of viral gene expression and the role of nucleosome positioning in the regulation thereof. The minimum of expression of the viral genome at 11-12 hpi correlates with a minimum in accessibility measured 12 hpi by ATAC-Seq. Furthermore, ATAC-Seq revealed a switch in the +1 nucleosome occupancy as early as 4 hpi.

3.4 Discussion

In this chapter, I have first employed 4sU-tagging of newly transcribed RNA followed by NGS to study the dynamic changes in transcriptional activity of NIH-3T3 fibroblasts during lytic mCMV infection. The increased sensitivity of nascent RNA-Seq towards on-going transcription compared to total RNA-Seq allowed me to identify host gene clusters with different expression kinetics that were associated with specific cellular functions. ATAC-Seq revealed a binary accessibility switch for gene promoters correlated with active or inactive transcription, respectively. Further, accessibility data was used to improve the *in silico* prediction of TFBS. Moreover, ATAC-Seq detailed out how nucleosome positioning within the viral genome changes over time after host infection.

First, RNA analysis by biosynthetic tagging with 4sU followed by sequencing enables sensitive and specific queries of how nascent transcription is regulated on a genome-wide scale.

Thio-substituted nucleotides do not occur naturally in living cells. Incorporation into nascent transcripts by RNA polymerases allows subsequent isolation of labelled RNA from bulk RNA. Hence, one can easily measure synthesis and decay rates in a given period, revealing dynamic changes in gene expression. Many host cellular transcripts are stable and mask the transcriptional response to infection. This is especially important when studying the early stages of mCMV infection, because of virion-associated RNA species, which are randomly incorporated into viral particles and thus resembling the cellular RNA profile of late stage infection. I successfully applied 4sU-Seq to comprehensively study newly transcribed RNA upon lytic mCMV infection. My libraries contain intronic sequences and other short-lived RNA species, such as promoter anti-sense transcripts (Figure 3.1a), and are strand specific (Figure 3.1a and b), what makes them a valuable unique resource to study sense and anti-sense transcription at viral and host, coding and non-coding sequences at the actual level of transcription. This allows a direct examination of the functional output of a gene upon viral infection. Furthermore, the transcriptional output can be correlated with regulatory inputs such as nucleosome positioning and TFBS availability. TU-tagging would be even possible *in vivo*, in a cell type specific manner (Gay et al., 2014), thus enabling the study of dynamic changes upon infection *in vivo* in a cell population of interest. This would provide new insights into how the observed transcriptional changes *in vitro* are also observed *in vivo*.

The provided data further our understanding of the transcriptional landscape of NIH-3T3 cells and, most importantly, of transcriptome changes upon viral infection: However, information on certain RNA species are limited. For example, RNA purification and library construction, select against short RNA species, such as miRNAs. Further, lowly expressed and/or very short-lived transcripts, such as enhancer RNAs (eRNAs) and promoter anti-sense RNAs (Rabani et al., 2014), exhibit accordingly low sequencing signals. Alternative methods, such as CAGE-Seq (Shiraki et al., 2003), NET-Seq (Churchman & Weissman, 2012) and TT-Seq (Schwalb et al., 2016) could be used to obtain a more elaborate picture on very short and transient transcripts. Especially the latter, TT-Seq, has been utilised to map eRNA and mRNA every 5 min after T cell stimulation with high sensitivity and identified numerous new primary response genes (Michel & Demel, 2017). TT-seq revealed in addition to mRNAs several hundreds of eRNAs changing significantly in expression within only 15 min after stimulation. This suggests that transcriptional activity in T cells changes immediately upon stimulation. Even though 4sU-Seq provides high sensitivity and temporal resolution, I could not detect many genes with a greater than 2-fold regulation between 0-1 hpi (Figure 3.2a). This might be due to the difference between fibroblasts and T-cells. The latter are primed towards fast immune responses.

In 3T3 cells I could observe mCMV induced expression patterns, including a rapid inflammatory response, an early transient induction of genes functionally associated with gene expression

and RNA metabolism as well as a transient peak of genes involved in proteolysis. All of these were counter-regulated later on. Notably, I could reproduce the previously described over-representation of YY1 binding sites in promoter regions, determined by the -500 bp to +100 bp definition, of genes with a peak at 3-4 hpi. After refining PPRs with ATAC-Seq, I found an over-representation of ETS- family TFBS, such as ELK1 and ELK4, which are activated by the MAPK/ERK pathway. This also had been described in human cells upon infection with hCMV (Caposio et al., 2010). Furthermore, refinement of TF site prediction using ATAC-Seq identified clusters potentially regulated by the proto-oncogene c-Myc and ETS family members, respectively. Notably, the over-representation of ETS family TFs in multiple clusters correlated with a strong induction at 5-6 hpi, which was followed by a reduction in expression levels 11-12 hpi. This suggests a more general role of ETS family TFs in the early events of mCMV infection. I did not identify ETS sites using genomic windows of fixed length around TSS, which highlights the importance of the incorporation of accessibility data into promoter analysis.

In this thesis, I provided potential host cellular TFs that might be involved in regulating host cellular gene expression, but I am lacking to explain what role viral proteins might play. The counter-regulation of the NF-κB response is consistent with the mCMV M45 gene product efficiently targeting NF-κB- signalling (Fliss et al., 2012). For hCMV it has been described that the viral protein IE1 increases p53 activity by phosphorylation through ATM, an important kinase in the DNA damage response (Castillo et al., 2005). I observe a delayed but constant increase in expression of genes involved in DNA repair (cluster 9). The mCMV ie1 gene, which is the orthologue of the hCMV IE1, shows traditional immediate early kinetics, resulting in a prominent peak of expression 1-2 hpi, suggesting against the direct positive regulation of cluster 9 by this viral TF. A previous study suggested the involvement of an unknown viral gene product in the counter-regulation of genes involved in the ER stress response they observed (Marcinowski et al., 2012). I observed a cluster with similar kinetics, containing genes involved in proteolysis (cluster 13). It is, therefore, tempting to speculate that a viral protein might be involved in the counter-regulation of cluster 13 as well. For hCMV the viral pUL38 protein is performing that function (Qian et al., 2011), whereas the mCMV protein remains unknown. The previous study on mCMV-infected cells also suggested that down regulation of two clusters required viral gene expression (Marcinowski et al., 2012). Taking together their findings, all cellular signalling pathways they identified to be induced during early mCMV infection, seem to be rapidly counter-regulated by the virus later on. Moreover, down-regulation of defined cellular signalling pathways prevails and thus most likely represents an intentional action of the virus to facilitate its own needs. I could identify two cellular gene clusters, which displayed persistent down-regulation, namely cluster 11 and cluster 3. Genes involved in chromatin modification as well as genes involved in cell proliferation and actin filament-based processes belonged to those clusters. Strikingly, within a few hours of infection, mCMV-infected cells

show a profound cytopathic effect. The underlying molecular mechanism is yet to be elucidated. It is tempting to speculate that transcriptional down-regulation of genes involved in actin filament-based processes and cell adhesion contributes to this phenomenon. Further, infected cells are arrested in either the G1 or the G2 stage of the cell cycle and hence stop proliferating. This cell cycle arrest seems to be mediated via cell cycle independent *ie3* gene expression (Wiebusch et al., 2008). How the expression of genes in those two clusters remains repressed at later stages of infection, when expression levels of the viral immediate early gene products drop, remains elusive.

Transcription only represents the first step in the expression of genes and many posttranscriptional regulatory layers can be envisaged to modulate expression levels. For example, it would be interesting to study which cellular, or viral, RNA-binding proteins are regulating transcript levels and transcript processing using CLIP procedures (Hafner et al., 2010; Ule et al., 2003). On the other hand, methods such as translating ribosome affinity purification (TRAP) (Heiman et al., 2014) or ribosome profiling (Ingolia et al., 2009), provide information on which transcripts are actually transported to the cytoplasm and are translated. Ribosome profiling data could provide a unique opportunity to quantify the rate of viral and host protein production throughout the course of infection. For example, it is still not clear if the transcripts from the early burst of gene expression upon herpes virus infection are actually translated. One way by which viruses influence host physiology is interaction with the cellular translation machinery. Not only do viruses depend on this machinery for their protein production, they also must block host defences, which inactivate the cellular translation apparatus. Recently, it has been reported that upon lytic HSV-1 infection transcription of many host genes progresses beyond poly(A) sites for several kilobases (Rutkowski et al., 2015). This in turn resulted in novel intergenic splicing events. Of note, hundreds of cellular genes seemed to be transcriptionally induced but were not translated. Thus, performing 4sU-Seq and ribosome-profiling measurements in parallel would allow direct assessment of differential translation of any cellular and viral genes during infection. Lastly, recent advances in mass spectrometry techniques enabled accurate measurements of changes in the human proteome during infection (Weekes et al., 2014). Since these measurements provide quantitative evaluations of steady-state protein levels during infection and ribosome profiling provides quantitative measurements that reflect the rate of protein synthesis, integration of these measurements could facilitate global identification of protein degradation rates during infection.

The genomic era began with Sanger sequencing of the bacterial DNA virus φX174 in 1977 (Sanger et al., 1977) and the mammalian DNA virus simian virus 40 (Fiers et al., 1978) the year later. However, even with the advent of next generation sequencing, the compact viral genomes with overlapping expression units makes it difficult to map all genes with

protein-coding potential. 4sU-Seq profiling of viral gene expression in combination with ribosome profiling and mass spectrometry offers an unprecedented opportunity to better annotate viruses. In addition, such methods can help to reveal novel aspects of the complex interaction between viruses and their hosts.

Traditionally, mCMV genes were categorised into immediate early, early and late genes, based on the time during infection when these transcripts could be detected in total RNA or by western blotting. This temporal regulation, however, does not seem to hold true when measuring levels of newly transcribed RNAs. An early burst of all viral classes of transcripts (immediate early, early and true late genes) had been reported before and was further shown to be dependent on the MOI, with a slight shift towards later time points for the early and late genes at a lower MOI (Marcinowski et al., 2012). Not only the induction of all classes of genes was surprising, but also the dramatic drop of transcriptional activity of all viral genes starting to occur 5-6 hpi and reaching an absolute minimum at 11-12 hpi. For some genes repression even continued throughout the entire infection. Only late genes showed an increase in transcription just after the onset of viral DNA replication. Strikingly, this can also be observed for host gene expression, with at least 4 host cellular gene clusters (cluster 2, 5, 8 and 13) showing a very clear and dramatic drop in expression at 11-12 hpi. Furthermore, the overall kinetics of viral gene expression throughout the entire infection best matches those of cellular genes within clusters 2, 5, 8 and 13. Notably, ETS family TFs and Ap-1 binding sites are characteristic for genes within those clusters. This suggests a major role of these TFs in regulation transcriptional events not only in the host but potentially also in the virus. Moreover, chromatin modifications of the viral genome are well known to play an important role during productive CMV infection (Sinclair, 2010). The ND10 body-associated protein Daxx is known to rapidly repress transcription of incoming viral genomes by inducing repressive chromatin modifications around the hCMV MIEP within 3 hpi. The peak of early gene expression should thus only occur after ND10 body-mediated repression has already been efficiently disrupted. Therefore, this suppression between 5-6 hpi and especially at 11-12 hpi, is unlikely to be due to the intrinsic antiviral defence mediated by ND10 bodies. It is rather striking, that during this stage of infection the viral genome displays the least overall accessibility reflecting tight association with histones genome-wide on the viral genome. This provides a platform for the decoration with repressive histone modifications, such as H3K27me3 put on by Polycomb group proteins, which have been shown to regulate herpes viral latency (Watson et al., 2013).

3.5 Conclusion

Taken together, the data presented in this chapter provides a comprehensive analysis of dynamic changes in gene expression upon lytic mCMV infection over time. This allowed for soft clustering of expressed genes into 14 groups with functional annotations, based on their

kinetics. The work presented here extends previous high-throughput studies. To the best of my knowledge, in this thesis I am the first to combine transcriptional profiling in real time, while simultaneously examining the proximal regulatory input, both genome-wide. This allowed me to reveal new potential roles of ETS TFs family in regulating transcriptional events occurring upon lytic mCMV infection in both, the host and the virus for the first time.

4 Structural changes of host and viral genome architecture upon lytic mCMV infection

4.1 Introduction

Regulation of gene expression extends far beyond the linear sequence and epigenetic control at PPRs. Genomes of higher eukaryotes do not exist as a one-dimensional polymer, or function in a sequential manner; rather, they are functioning in a 3D space, the cell nucleus. Recent work has led to an increasing body of evidence that the genome structure relates to genome function and that gene expression can be regulated by distal regulatory elements (Bulger et al., 2011), which can be located up to several Mb away (Bickmore, 2013; Sanyal et al., 2012). For those distal regulatory regions to act on transcription of genes located far away on the linear sequence, and frequently with several unaffected genes in between, they have to be in close physical proximity to the promoters they regulate (Amano et al., 2009). Lytic mCMV infection is known to be associated with dramatic changes in host 3D nuclear architecture, due to immensely growing VRCs. Additionally, more than 40 % of all genes show a >2-fold differential expression at 47-48 hpi (Figure 3.2a). This makes lytic mCMV infection a suitable model system to study not only proximal (Chapter 3) but also distal gene regulation and the contribution of the overall nuclear organisation.

The nucleus of the interphase eukaryotic cell is a highly compartmentalized structure containing the three-dimensional network of chromatin and numerous proteinaceous subcompartments. DNA viruses induce profound changes in the intranuclear structures of their host cells. Early upon infection, the incoming viral genomes can be found juxtaposed to the PML-ND10 nuclear bodies. The PML-associated cellular transcriptional repressor Daxx inhibits hCMV gene expression from the MIEP, which is circumvented by the viral tegument protein pp71. This was thought to be part of the anti-viral cellular defense. However, it has now been shown for PRV, a neuroinvasive swine alphaherpesvirus, that despite the large number of viral genomes entering the nucleus, fewer than seven incoming genomes may actually be actively transcribed and later on replicated. This led to the speculation that PML bodies are involved in the initiation of viral replication (Kobiler et al., 2011; Kobiler et al., 2010). Viral DNA replication in the nucleus is a highly structured process in which, in the few well-documented cases, large discrete VRCs are formed. Previous studies have provided evidence for a number of morphological changes that take place in the nucleus during herpes viral infection, including the margination of host chromatin and disaggregation of the nucleolus (Schwartz & Roizman, 1969; Sirtori & Bosisio-Bestetti, 1967). Additionally, it is known that the VRC excludes host chromatin (Puvion-Dutilleul & Besse, 1994) and that margination of host chromatin takes place without the loss thereof, which both led to the conclusion that the profound spatial

rearrangement of the infected cell`s genome occurs concomitantly with the expansion of the VRCs (Monier et al., 2000). Markedly, despite the increase in viral DNA content of the nucleus, a significant increase in the protein mobility was observed in infected compared to non-infected cells (Ihalainen et al., 2009). Early during herpesvirus infection, the interchromatin domains enlarge and the nuclear volume increases as much as twofold (Monier et al., 2000). At the same time, VRCs expand, move, and coalesce, but do not mix. A model in which herpesvirus infection increases the porosity of the nucleus had been proposed (Bosse et al., 2015). Diffusion of proteins in enlarged DNA corrals allows the majority of viral capsids, which are large protein aggregates, to have direct access to the nuclear envelope, which is necessary for viral nuclear egress. Furthermore, margination of host chromatin brings it closer to the nuclear periphery and closer to the nuclear lamina, an environment that helps to establish and maintain interphase chromosome topology and thus the overall genome spatial organization. Furthermore, lamina-associated chromatin domains (LADs) have been described as gene poor, late replicating and generally inactive and compacted (Guelen et al., 2008). A variety of DNA-binding proteins may anchor specific loci to the nuclear lamina, while histone-modifying enzymes partly mediate the local repressive and compacting effect of the lamina (Kind & van Steensel, 2010). Several lines of evidence indicate that the major driver of chromatin compaction are Polycomb group proteins, especially the Polycomb orthologue CBX2 in the mouse (Grau et al., 2011).

4.2 Objectives and outline

The results presented in chapter 3 demonstrate the immense, genome-wide changes in transcriptional activity with the ongoing mCMV infection and shed some light on how DNA accessibility at PPRs is involved in gene regulation. Furthermore, *in silico* TFBS prediction offered some first promising TF candidates that might play a role in mediating these observed changes. However, these data only allow for linear description of the genome, and do not match genes with distal regulatory elements, nor do they present an overview of global architectural changes upon lytic mCMV infection. In order to detect host cellular and viral changes in genome organisation and to understand how these changes correlate with the observed transcriptional alterations, I performed Hi-C and capture Hi-C experiments on lytically mCMV infected NIH-3T3 fibroblasts. These libraries contain the information of the 3D organization not only of the host, but also of the viral genomes and the spatial correlation between both. Since Hi-C experiments are cost intensive, I decided to focus on three time points during infection: non-infected cells, 2 hpi and 48 hpi. During the early time point of infection, namely 2 hpi, the first host cellular genes are dramatically induced (cluster 6, Figure 3.4) and the early burst of viral gene expression is starting to occur (Figure 3.13a). Since it had been reported that this early burst of viral gene expression shifts to 3-4 hpi when using an MOI

= 0.5 (Marcinowski et al., 2012), I included this low MOI early time point as well. In order to document the dramatic re-arrangement of host cellular nuclear architecture, I included the late stage time point of infection (48 hpi).

In the first part of the chapter, I will focus on the broader picture of genome architecture and major changes thereof occurring post-infection. I will demonstrate how the overall folding pattern of the host cellular genome is affected by viral infection. Further, I will correlate chromatin and functional properties of Mb-scale domains with their transcriptional output at the different stages of infection. Understanding these changes in chromatin architecture and chromosomal territories during CMV infection is likely to provide important insight not only into the biology of infection but also into the complex relationships between chromatin structure, gene activity, and the functional state of the cell in general. Hi-C and capture Hi-C will further strengthen our understanding of how the viral genome is arranged in a 3D manner and reveal host-pathogen DNA-DNA interactions. So far, little is known about which intra-nuclear areas the viral genomes localize to, which factors determine whether they are actively transcribed or become latent, or how the viral genome itself is arranged in a 3D manner. Furthermore, it is unclear whether additional enhancer-like sequences, in addition to the known bipartite major immediate-early transcriptional enhancer (Kropp et al., 2009), exist in the genome. Additionally, the association and interaction of the VRCs with the cellular chromatin at later stages in infection is poorly characterized.

In the second part of this chapter, I will explain how we can use capture Hi-C data to examine regions of interest, determined by substantial differential expression in the 4sU data, at high resolution. This will reveal that contacts between promoters and distal *cis*-regulatory regions are stable during the course of viral infection.

4.3 Results

4.3.1 Hi-C library preparation and quality controls

I prepared Hi-C libraries from NIH-3T3, lytically infected with BAC-derived mCMV Smith strain, in order to generate 3D contact maps from the same cells as the RNA-Seq and ATAC-Seq datasets. This also matched the high quality, previously prepared replicate of Hi-C libraries from non-infected NIH-3T3, 2 hpi and 48 hpi using an MOI = 10, and the 4 hpi time point using an MOI = 0.5, generated as part of my master project (referred to as replicate I). The major steps of the Hi-C protocol are outlined in Figure 4.1a. The Hi-C libraries analysed in this chapter have been prepared using the “in-solution” ligation during Hi-C library preparation (Schoenfelder et al., 2015a), whereby the nuclei are lysed prior to ligation and the ligation step is carried out in a diluted fashion in order to favour ligation between cross-linked fragments.

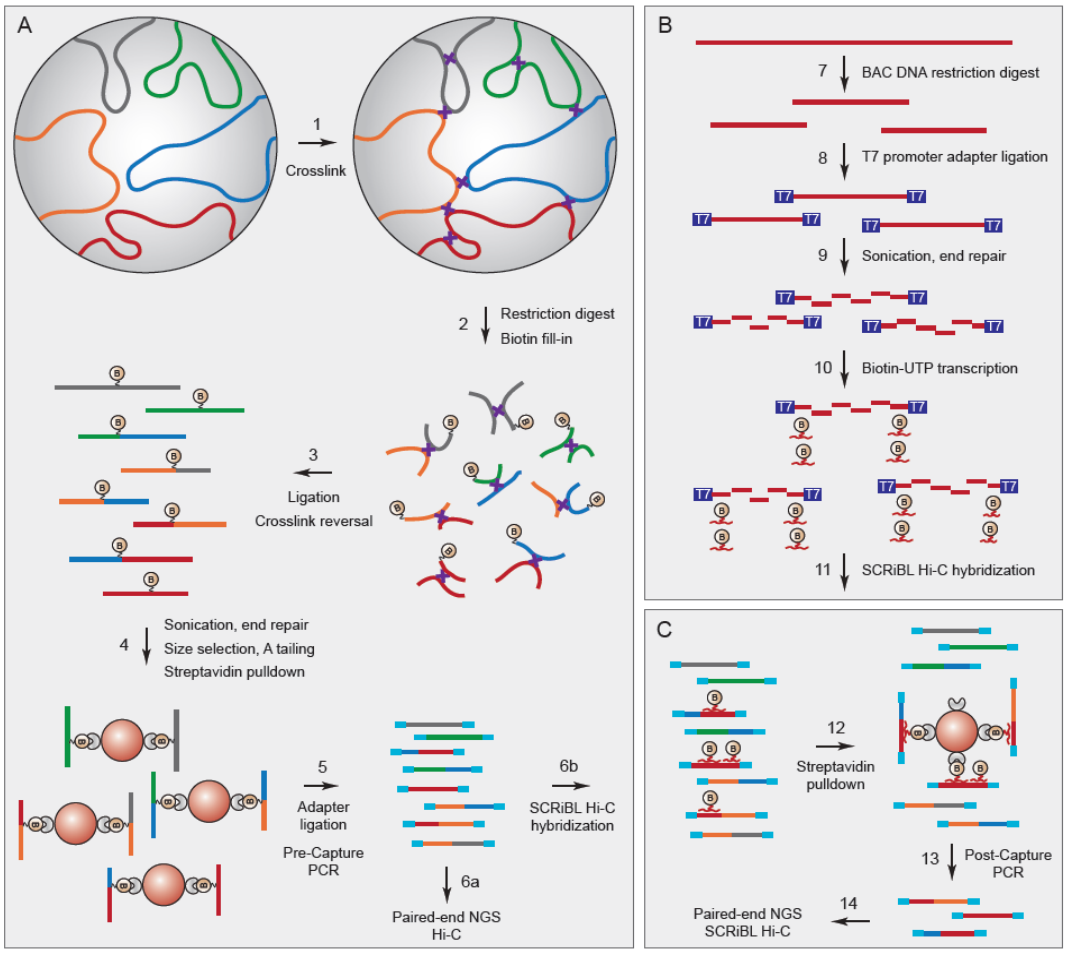


Figure 4.1 | Schematic of Hi-C and SCRiBL library preparation

(a) Schematic summary of the main steps involved in Hi-C library generation. In brief, chromatin contacts are preserved using formaldehyde, chromatin is digested using a restriction enzyme and restriction fragment ends are filled in using biotinylated nucleotides. After blunt-end ligation of fragment ends in close proximity, crosslinks are reversed and libraries are subjected to sonication, end repair, a double-sided size selection and A-tailing. A biotin-streptavidin pull-down enriches for successful ligation events. After adapter ligation and PCR amplification, libraries can be subjected to SCRiBL or NGS. **(b)** SCRiBL biotinylated RNA bait generation is based on BAC DNA covering the genomic region of interest. DNA is digested using the same restriction enzyme as for the Hi-C library generation, followed by T7 promoter ligation. Sonication and subsequent end-repair are followed by in vitro transcription using biotinylated rUTP nucleotides. **(c)** Hi-C libraries are hybridised with complementary biotinylated RNA baits at 65°C for 24 h. A subsequent streptavidin pull down enriches for Hi-C di-tags of interest, which are PCR amplified and sequenced. This figure was generated by Stefan Schoenfelder (unpublished).

Following purification of ligated DNA (Figure 4.1a, step 3), I performed ligation efficiency control PCRs to confirm the presence of a specific Hi-C ligation product. This ligation product, between two histone genes, *Hist1h4i* and *Hist1h4f*, was selected based on previous mouse Hi-C and 3C PCR results (Biola-Maria Javierre, Mayra Furlan-Magaril, Stefan Schoenfelder, personal communication) and spans more than 1.5 Mb on the linear sequence (Figure 4.2a). Further, a short-range 3C contact, between the calreticulin TSS and one of its introns ~20 kb away on the linear sequence, was amplified using two forward primers, which would not give

an amplicon on the linear sequence. Restriction digestion and biotin fill-in prior to ligation create a new ClaI restriction site at the ligation junction and the BglII one is lost (Figure 4.2b). I, thus, confirmed that the majority of the amplified ligation products contain the correct junctions (Figure 4.2c). After enriching for successful ligation events with a biotin-streptavidin pull-down and the addition of sequencing adapters, I estimated the number of PCR cycles needed to amplify approximately 1 µg of final Hi-C library, which is sufficient for subsequent capture and sequencing of the library. This was done by running test PCR amplifications on aliquots (1/20th library) with different amplification cycles and visualisation of the products following agarose gel electrophoresis (Figure 4.2d). The aim of these tests is to avoid over-amplifying the libraries, which can lead to PCR duplication artefacts. Previous work has indicated that Hi-C libraries, generated from 20 million cells starting material, should be amplified using a number of PCR cycles at which a smear is faint but clearly visible (Stefan Schoenfelder, personal communication). All of the libraries were amplified using 8 cycles.

After amplification and purification, the size distribution and the concentration of the Hi-C libraries were assessed by Bioanalyzer analysis (Figure. 4.2e). Libraries were found to be in the range between 300-700 bp, which is optimal for Illumina sequencing, demonstrating precise sonication and size selection. All four libraries showed sufficient but not over amplification with concentrations between 30 nM to 95 nM. These concentrations were confirmed by Kapa qPCR. All Hi-C libraries were sequenced on one HiSeq 2500 lane with a 50 bp paired-end setup.

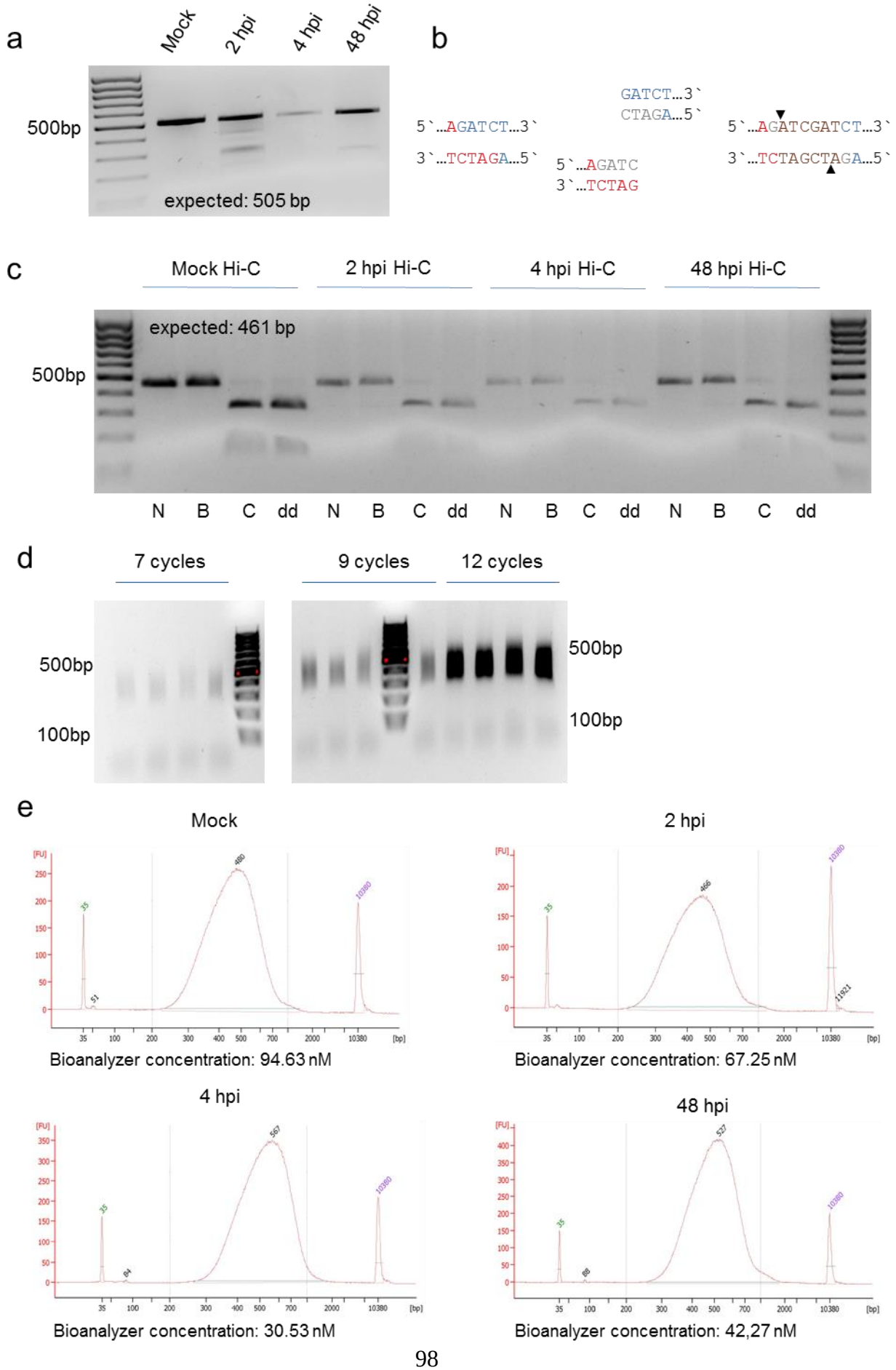

a
Mock
2 hpi
4 hpi
48 hpi
500bp
expected: 505 bp
b
5`...AGATCT...3`
3`...TCTAGA...5`
GATCT...3`
CTAGA...5`
5`...AGATC
3`...TCTAG
5`...AGATCGATCT...3`
3`...TCTAGCTAGA...5`
c
Mock Hi-C
2 hpi Hi-C
4 hpi Hi-C
48 hpi Hi-C
500bp
expected: 461 bp
N B C dd N B C dd N B C dd N B C dd
d
7 cycles
9 cycles
12 cycles
500bp
100bp
500bp
100bp
e
Mock
Bioanalyzer concentration: 94.63 nM
2 hpi
Bioanalyzer concentration: 67.25 nM
4 hpi
Bioanalyzer concentration: 30.53 nM
48 hpi
Bioanalyzer concentration: 42,27 nM

Figure 4.2 | Quality control test during Hi-C library preparation

(a) Ligation test using PCR to amplify a known Hi-C ligation product between the two histone genes, *Hist1h4i* and *Hist1h4f*, followed by agarose gel electrophoresis and visualisation. The expected size was 505 bp .Primers were previously shown not to give an amplicon on genomic DNA. **(b)** Schematic of how restriction enzyme digestion with BglII (blue and red) followed by fill-in (grey) and blunt-end ligation creates a ClaI restriction site (brown), while the BglII site is lost. **(c)** A short range contact between the calreticulin TSS and the first intron, ~20 kb away on the linear sequence, was amplified using PCR and was either not digested, or digested with BglII, ClaI or both (as indicated) to test the identity of ligation junctions. Desired junctions can only be digested using ClaI but not with BglII. **(d)** PCR test were performed on 1/20th of the library material to determine the optimal number of PCR cycles for the final PCR amplification. The desired amount of final library for subsequent capture Hi-C and NGS is around 1 μg. **(e)** Bioanlayzer profiles for Hi-C libraries following PCR amplification and purification. Library concentrations are reported for the range of 200-1000 bp and were confirmed by Kapa qPCR.

Hi-C has the potential of capturing the ensemble of chromosomal interactions within a cell population, but one caveat of this approach is the vast complexity of mammalian Hi-C libraries, estimated to contain 10^{11} unique pair-wise interactions (Belton et al., 2012). This impedes their analysis at a resolution needed to identify specific promoter to enhancer interactions. To overcome this limitation, protocols including sequence capture steps have been proposed (Dryden et al., 2014; Hughes et al., 2014; Mifsud et al., 2015; Schoenfelder et al., 2015a), all of which have in common that they require expensive biotinylated custom made probes. To circumvent this, I generated biotinylated RNA probes, covering solely the ends of restriction fragments of 5.6 Mb, from BACs, ranging from 150-250 kb, and enriched my Hi-C libraries for regions of interest, defined by dramatic transcriptional changes (see Table 2.3 for the BACs used and the gene of interest). A schematic of the protocol is outlined in Figure 4.1b. In brief, BAC DNA is isolated and digested with the same restriction enzyme that was used to generate the Hi-C libraries. T7 promoter adapters are ligated to the restriction fragment ends, enabling *in vitro* transcription using biotinylated rUTPs after sonication. This results in biotinylated RNA probes complimentary to the restriction fragment ends present in the Hi-C library. BAC DNA for 24 regions of interest (five control genes, six continuously down regulated, seven continuously upregulated, three Nf-κB regulated and three IF regulated genes) was pooled and kept separate from BAC DNA covering the mCMV genome, to enable separate enrichment for host genomic regions and the viral genome. Previous experiments have shown that there is an estimate of 50 viral copies present at 2 hpi, which corresponds to a 25 fold excess of the viral genome compared to a similar sized locus of the diploid host genome, hence the 1:24 ratio. Efficient sonication after ligation was verified by gel electrophoresis of sonicated and non-sonicated BAC DNA (Figure 4.3a). Furthermore, *in vitro* transcription produced RNA in the desired size range, centred on 180 nt (Figure 4.3b), which was biotinylated (Figure 4.3c). The generated Hi-C libraries were enriched using these biotinylated RNA baits, as described under 2.11, in separate reactions for the host and the viral baits. Additionally the pre-existing high

quality replicate of Hi-C libraries was enriched using these biotinylated RNA oligonucleotides. Test PCRs were performed in order to determine the number of cycles needed to obtain sufficient material for NGS, but not to over-amplify the libraries (Figure 4.3d). Following 8 cycles of PCR amplification and library purification, all libraries showed a desired concentration and size distribution, determined by Bioanalyzer analysis (Figure 4.3e). These concentrations were verified by Kapa qPCR and the enriched Hi-C libraries were subjected to next generation sequencing on the HiSeq 2500 platform, using a 50 bp paired-end output.

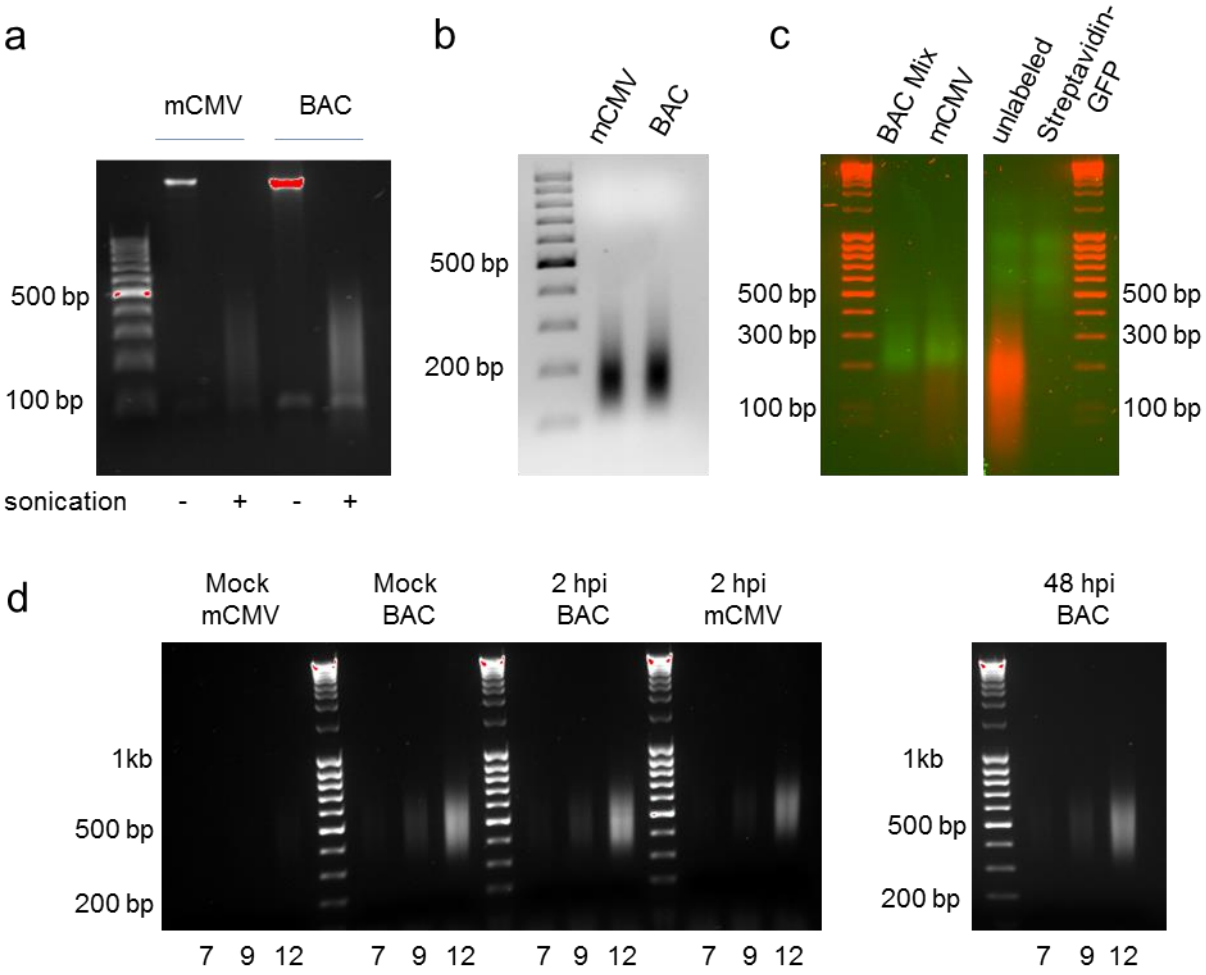

a
mCMV
BAC
500 bp
100 bp
sonication - + - +
b
mCMV
BAC
500 bp
200 bp
c
BAC Mix
mCMV
unlabeled
Streptavidin-GFP
500 bp
300 bp
100 bp
d
Mock mCMV
Mock BAC
2 hpi BAC
2 hpi mCMV
48 hpi BAC
1kb
500 bp
200 bp
7 9 12
e
Mock BAC
Bioanalyzer concentration: 17.15 nM
2 hpi BAC
Bioanalyzer concentration: 14.23 nM
2 hpi mCMV
Bioanalyzer concentration: 14.45 nM
48 hpi BAC
Bioanalyzer concentration: 15.60 nM

Figure 4.3 | Quality control test during SCRiBL bait and SCRiBL library preparation

(a) BAC DNA, for the mCMV genome alone (left two lanes) and for 24 BAC combined (right two lanes) was subjected to restriction enzyme digest using BglII**,** followed by BglII overhang specificT7 promoter adapter ligation. Ligation products prior to and post sonication were separated by agarose gel electrophoresis and visualised, indicating successful sonication to a desired size range between 100-500 bp. **(b)** Sonicated DNA was *in vitro* transcribed using biotinylated rUTP. RNA was purified using the T7 MegaClear columns and 1 μg each, of the mCMV and the BAC RNA mix, was separated according to size by agarose gel electrophoresis. RNA oligomers showed the expected size of 100-300 nt. Note that the biotin moiety can influence RNA mobility during gel electrophoresis. **(c)** 250 ng per RNA sample were incubated with GFP-tagged streptavidin 647 (diluted 1:200) and agarose gel electrophoresis on a 2 % gel was performed. RNA was stained with gel red and scanned with the Typhoon. Red and green channel pictures were overlaid. In both negative controls (unlabeled baits and only streptavidin 647) the streptavidin ran as a high molecular smear. In all samples this smear shifted completely to lower molecular weights and ran directly above the RNA, thus indicating interaction between the two. **(d)** Test PCR were performed on a 1/20th of the SCRiBL libraries to determine the optimal number of final PCR amplification cycles, in order to obtain just enough material for NGS. **(e)** Library concentration and size distribution were assessed by Bioanalyzer analysis. Libraries showed the expected size range and concentrations. Given concentrations are measured for the range of 200-1000 bp and were confirmed by Kapa qPCR.

After completion of each of the sequencing runs, sequencing reads were processed using HiCUP (Wingett et al., 2015) to align reads and remove experimental artefacts and PCR duplicates. An additional Perl script was run on the SCRiBL libraries to remove off-target read-pairs. A schematic of Hi-C protocol intrinsic possible artefacts is depicted in Figure 4.4a. The exact output values from each step of the HiCUP pipeline can be taken from supplementary Table 1. In summary, 35-75x10^6 total pairs were derived in the respective Hi-C libraries. These numbers decreased to around 30 million for all libraries after filtering, which corresponds to 40, 50 and 80 % valid reads for 2 hpi, 48 hpi and the mock-infected Hi-C libraries, respectively (Figure 4.4b). These values increase to more than 70 % for all there libraries with the additional capture of host genomic Hi-C di-tags. Only when capturing the viral genome, 2 hpi, the percentage of valid reads decreased dramatically. Similar results were observed for the capture statistics of replicate 1. A good indicator of the libraries` quality is the *cis*/*trans* ratio within the valid pairs. Random ligation would result in 95 % *trans* (interchromosomal) reads, because there are 19 mouse autosomes (plus X and Y chromosomes), and thus the chance of two randomly ligated restriction fragments to originate from the same chromosome is around 1:20 (5 %). This has been experimentally verified (Mifsud et al., 2017). In contrast, in a normal Hi-C library, the physical linkage of sequences on the same chromosome imposes a strong bias towards *cis* interactions. It has been reported for “in-solution” ligation Hi-C experiments, that typically 55 % of the aligned reads represent interchromosomal (*trans*) interactions, while 45 % of the reads come from intrachromosomal (*cis*) interactions (van Berkum et al., 2010), although these percentages may vary, depending on the cell type analyzed. Previously described values were matched by the NIH-3T3 libraries,

whereas the 2 hpi libraries were slightly lower and the 48 hpi libraries were slightly higher in numbers of *cis* reads.

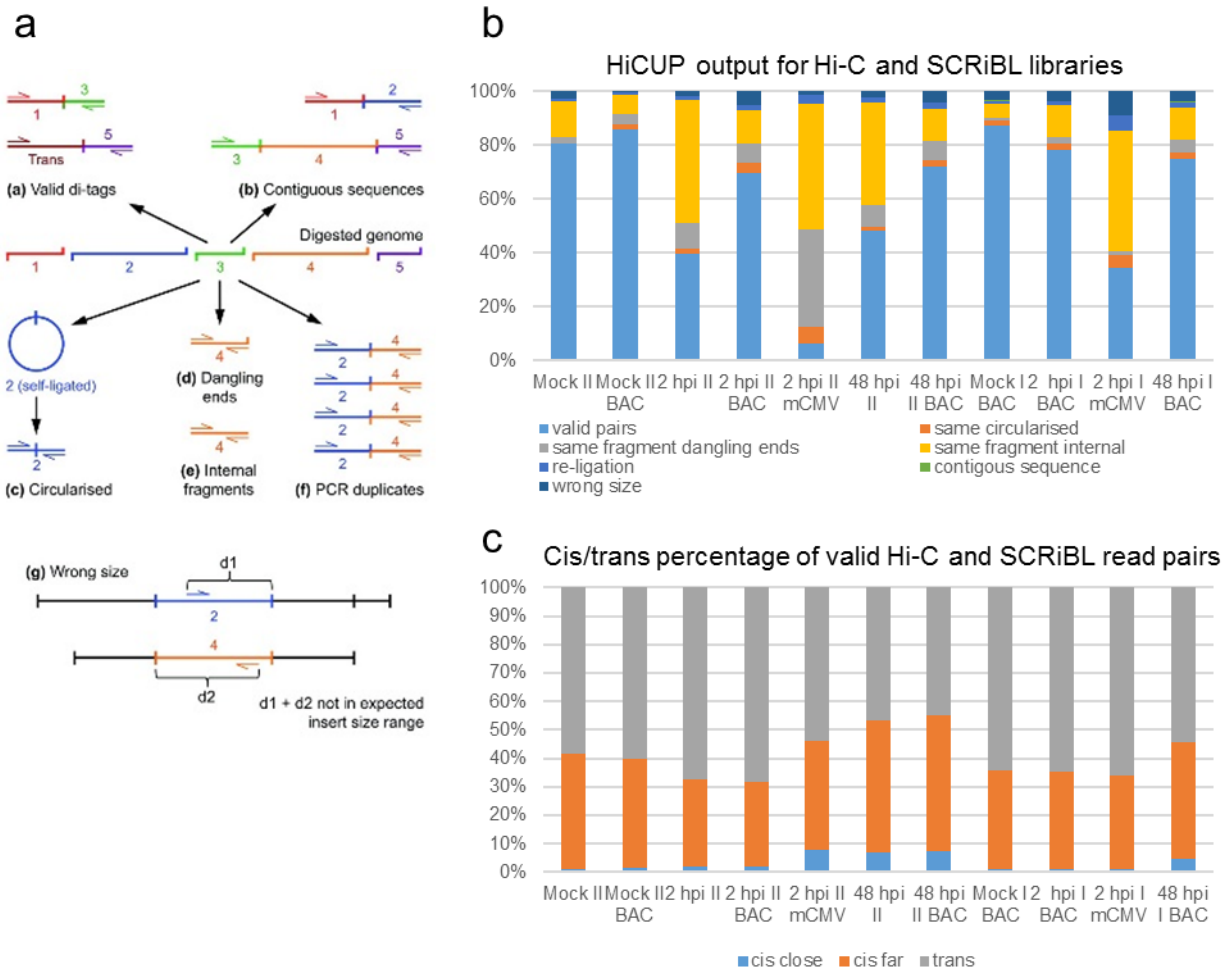


Figure 4.4 | Post sequencing quality control of Hi-C and SCRiBL libraries

(a) Schematic showing possible Hi-C protocol inherent artefacts. Only Hi-C di-tags where both ends are originating from restriction fragment ends coming from different fragments contain useful information about the 3D organisation, the rest are filtered out by the HiCUP pipeline (Wingett et al., 2015). **(b)** 100 % stacked column chart showing the percentage of valid and invalid read pairs, compared to the total number of mapped reads per library. Successful Hi-C libraries contain a high proportion of valid reads, usually higher than 70 %. **(c)** 100 % stacked bar chart displaying the percentage of close cis (<10 kb), far cis (>10 kb) and *trans* reads for the valid read-pairs in each of the generated Hi-C and SCRiBL libraries. 55 % trans reads have been reported for successful libraries using the "in-solution" ligation protocol (van Berkum et al., 2010). Absolute values can be taken from supplementary Table 1.

4.3.2 Global changes of the host cellular genome organisation

First, to verify that host nuclear architecture is globally disrupted under the experimental conditions used here, I stained non-infected cells and cells infected with mCMV at an MOI = 10 with DAPI and an anti-actin antibody coupled with Alexa Fluor (ab206277) and performed confocal fluorescence microscopy. Clear compaction of the host genome can be observed,

since apparent heterochromatic speckles in non-infected NIH-3T3 seem to merge 48 hpi into larger dense heterochromatic speckles, while large regions of only faint DAPI staining are emerging (Figure 4.5a), which most likely represent the VRCs. Actin staining revealed the known cytopathic effect (Gandhi & Khanna, 2004). Infected cells lose their typical fibroblast morphology and round up, while the overall size of the DAPI stained nucleus seems to marginally increase upon infection. Further indicative of successful infection is the vast number of Hi-C reads mapping to the viral genome 48 hpi, with nearly half of all valid reads mapping to the viral genome late in infection. This would result in almost 10,000 copies of the viral genome making up roughly half of the DNA inside the nucleus, without excessive expansion of the nucleus in size (Figure 4.5a) (Gibbs et al., 2013).

4.3.2.1 The local and global folding pattern of the host genome does not change dramatically upon lytic mCMV infection

Interaction data, obtained from Hi-C experiments, are usually represented in heatmaps, where one chromosome is plotted against itself, or genome-wide against all others. Colour coding reflects the enrichment of observed contacts between two genomic coordinates with respect to an expected rate of interactions between the same two loci based on a probabilistic model grounded on declining ligation frequencies with increasing genomic distance, which further is corrected for Hi-C method intrinsic biases (Imakaev et al., 2012). To assess the reproducibility between both replicates, I plotted corrected genome-wide heatmaps with 1 Mb resolution for the uninfected samples (Figure 4.5b) and 48 hpi (Figure 4.5c), where I expected the most dramatic differences. The overall appearance between the two replicates is very similar, with a strong signal of intrachromosomal (*cis*) interactions along the diagonal and mainly only diffuse interchromosomal (*trans*) interactions. This observation holds true for both time points. Notably, one striking trans interaction, amongst quite a few weaker ones, can be observed between a large region on chromosome 4 and chromosome 16 in all libraries, which is agreement with the most recently published karyotype of NIH-3T3 (Figure 4.5d), representing the largest translocation between two trisome chromosomes in a near-tetraploid genome (Leibiger et al., 2013). Surprisingly, there were no major differences between the contact maps of non-infected cells and the contact maps of the cells 48 hpi, despite the massive host genome re-arrangement observed by microscopy (Figure 4.5a) (Gibbs et al., 2013). Of note, for the 4 hpi time point an MOI = 0.5 was used, which resulted in about only a quarter of the cells being infected. Hi-C experiments are usually performed on large, ideally homogenous, cell populations, and only very strong effects originating from the smaller sub-population will be detected. Not surprisingly, the host cellular heatmaps at 4 hpi with the low MOI remarkably resembled the ones from non-infected NIH-3T3 at all resolutions analysed and were therefore excluded from further analysis in this part of the thesis. To further investigate changes in chromosome folding in more detail, I plotted heatmaps for individual chromosomes for both

replicates and all three time points (exemplary heatmaps for chromosome 15 are depicted in Figure 4.5e) and analysed the Pearson`s correlation coefficient between the pair-wise interactions. This further illustrates the high similarity between the two replicates but also between all three time points of infection.

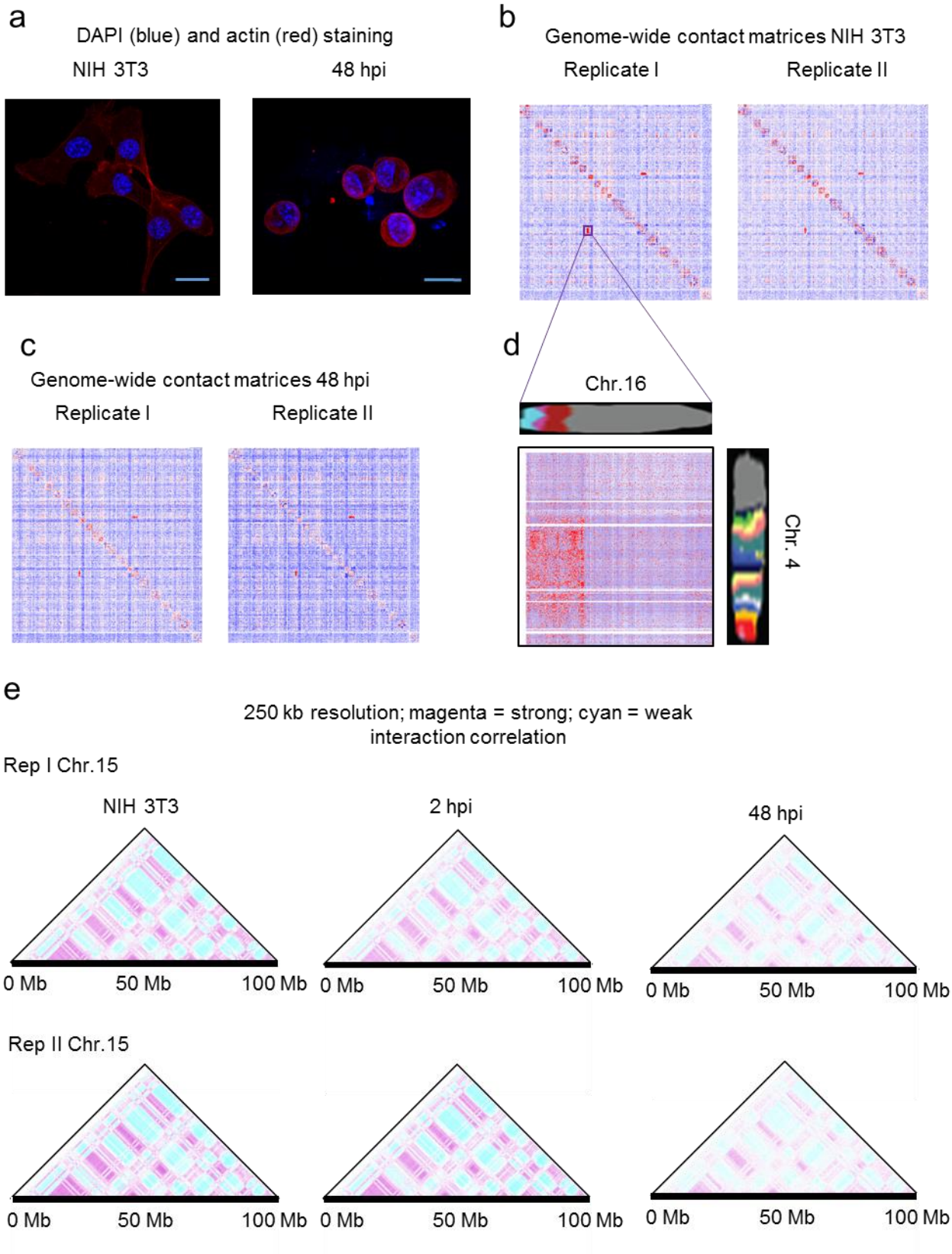


Figure 4.5 | Hi-C is capable of detecting the structural changes upon lytic mCMV infection

(a) Non-infected NIH-3T3 and NIH-3T3 48 hpi were fixed in formaldehyde and stained using DAPI (blue) and an anti-actin antibody coupled with Alexa Fluor (ab206277, red signal). On a Zeiss 780 confocal microscope 1 μm

z-stacks were imaged and merged using the Imaris program. Scale bar = 20 µm. Genome-wide Hi-C contact maps at 1 Mb resolution for both replicates shown for **(b)** non-infected NIH-3T3 and **(c)** NIH-3T3 48 hpi (MOI = 10). Chromosomes are arranged from left to right and top to bottom as follows: Chromosome 1,11-19, 2-9, X and Y. Read counts were normalised for sequencing depth, distance and Hi-C specific biases using an iterative approach (Imakaev et al., 2012). Red colouring indicates strong interactions, whereas blue is indicative of weak interactions. A strong red diagonal indicated the preference for intrachromosomal over interchromosomal interactions. **(d)** The one prominent interchromosomal interaction matches the previously reported translocation between chromosome 4 and chromosome 16 (Leibiger et al., 2013). **(e)** Corrected contact matrices at 250 kb resolution for chromosome 15. Shown is the Pearson`s correlation illustrating the correlation between the interaction profiles of every pair of 1 Mb loci along the chromosome (magenta = 1, cyan = -1). Displayed are both replicates for the three homogenous time points. The plaid pattern indicates the presence of two compartments within the chromosome. Interaction profiles are remarkably stable between replicates, but also between the different time points.

The overall interaction pattern of all chromosomes does not change dramatically with the ongoing infection, although the chromosome-wide folding pattern does seem to be less pronounced at the late stage of infection. Since on a more global level at 250 kb resolution, no changes could be detected, I wanted to increase the resolution of my Hi-C analysis to 50 kb and therefore, decided to pool the two replicates to increase coverage and the statistical power to detect interactions. Visual inspection of the data revealed a genome-wide compaction of some but not all self-interacting Mb-sized regions, accompanied by loss of interactions between these self-interacting domains. These regions seem to overlap with previously published TADs (Dixon et al., 2012). A heatmap of a representative region on chromosome 16 is displayed in Figure 5.6a. Overlaying the interaction data with publicly available ChIP-Seq data for H3K9me3 obtained from primary MEFs (Dixon et al., 2012), association with the nuclear lamina in MEFs (Peric-Hupkes et al., 2010) and my ATAC-Seq data, followed by visual inspection suggested that especially inactive stretches of DNA become more compacted and self-interacting (Figure 5.6b). To determine the genome-wide impact of this, I calculated the genome-wide pair-wise distance between all 10 kb bins and plotted it against the ligation frequency (Figure 5.6c). A strikingly higher frequency of short to mid-range interaction is detectable at 48 hpi, whereas 2 hpi a minor depletion of short range interactions compared to non-infected NIH-3T3 was observed. Those two findings are in good agreement with microscopy studies (Gibbs et al., 2013). Furthermore, separating the genome into inactive, LADs and active, inter-LADs and calculating the genome-wide distance decay revealed that both active and inactive regions are being compacted late in infection, but that this effect is more pronounced for the already inactive LADs.

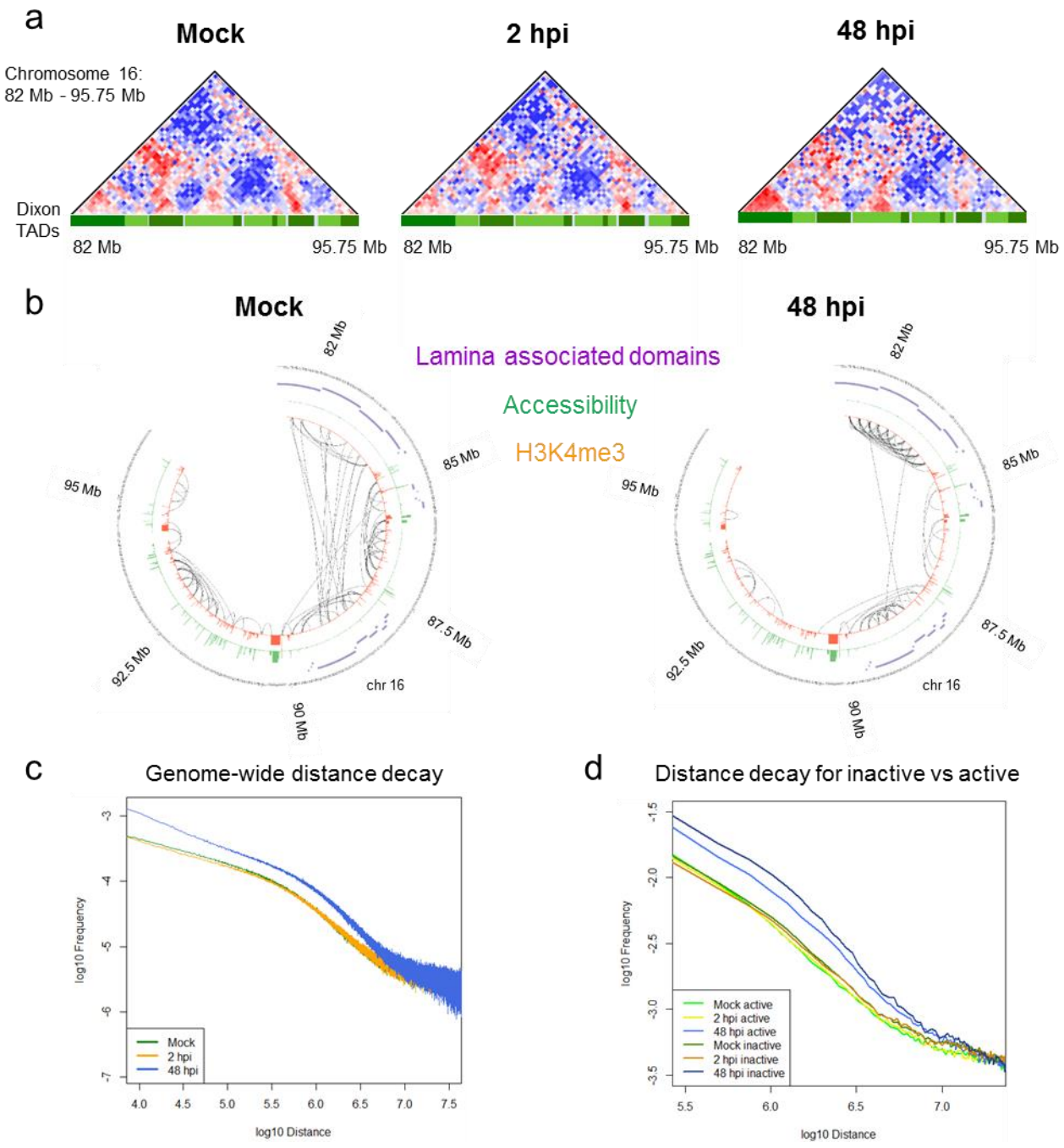


Figure 4.6 | Hi-C detects genome-wide compaction of inactive loci upon lytic mCMV infection

(a) Interaction profiles at 50 kb resolution for a 13.75 Mb long region on chromosome 16 for non-infected (left), infected cells 2 hpi (middle) and 48 hpi (right) from combined replicates, corrected for coverage and distance decay. Strong interactions are depicted in red, whereas weak interactions are shown in blue. TADs described in (Dixon et al., 2012) are alternatingly depicted underneath in different shades of green. Grey regions represent inter domain regions. Some TADs show strong compaction at 48 hpi, resulting in more intra-TAD interactions with specific loss of inter-TAD interactions **(b)** This phenomenon can also be visualized using circus plots, where the linear stretch of DNA is displayed as a nearly circular object and significant interactions, after distance and coverage correction, are displayed as arches. Additional information about association with the nuclear lamina (purple) and histone H3K4me3 (orange) are given and show that inactive regions are being compacted at this locus. **(c)** Genome-wide intrachromosomal contact probabilities between 10 kb stretches of DNA decrease as a function of genomic distance

for all the stages of infection. A clear increased probability for short to mid-range interactions can be observed for the 48 hpi time point. **(d)** Genome-wide 10 kb bins were separated based on their association with the nuclear lamina and intrachromosomal contact probabilities were plotted as a function of genomic distance for both categories in all three time points. This indicates that both categories are enriched for short distance contacts 48 hpi, although the inactive lamina associated domains show a more pronounced effect.

These results indicate that the overall folding pattern of the host genome does not change as dramatically as anticipated by microscopy studies. Nevertheless, I can observe a genome-wide compaction of the host genome late in infection, determined by an increase in local interactions, especially pronounced at already inactive regions near the nuclear periphery.

4.3.2.2 TADs are stable units of the genome, but are subjected to compaction at late stages of mCMV infection

At a resolution of 50 kb, highly self-interacting regions emerge on the diagonal of the heatmaps, seen as “triangles”. These regions were termed “topological associated domains” (TADs) and were found to be bounded by narrow segments, TAD-boundaries, which possess insulating properties (Dixon et al., 2012). With increasing resolution sub-domains were described to emerge until at very high resolution (4-cutter Hi-C) loops between 1 kb windows are visible (Rao et al., 2014). For the purpose of this thesis, I calculated TADs based on the directionality index, first described by Dixon et al. 2012. This resulted in around 1,800 detected TADs in non-infected cells and increased to ~2,000 detected TADs 2 hpi (Figure 4.7a). I could identify 2,200 TADs genome-wide in the 48 hpi libraries, which correlated well with the previously reported number of those domains (Dixon et al., 2012). Furthermore, the length of the described domains in all time points matched the previously reported size distribution, apart from long domains, which were not annotated as domains in my data (Figure 4.7b). This might explain the slightly reduced genomic coverage of ~60 % compared to the initially reported ~80 % (Figure 4.7c). TADs have been reported to be largely cell type invariant and are even conserved amongst species (Dixon et al., 2012). Additionally, changes in transcriptional programs or cellular identity, predominantly lead to changes in interactions occurring between elements in the same TAD without any major alterations in TAD boundaries (Dixon et al., 2015; Freire-Pritchett et al., 2017). Although changes in 3D genome architecture that involve TAD boundaries have been reported. A good example of a change in TAD boundary detection occurs during embryonic development within the murine *Hox* D cluster (Andrey et al., 2013). To detect and document any changes in TAD boundary formation upon lytic mCMV infection, I calculated the overlap of TAD boundaries between the different stages of infection. Modest overlap with ~1,400 TAD boundaries detected at all three stages of infection was observed. Furthermore, an increasing amount of domain boundaries, unique to the individual time points, was observed with the ongoing infection, but overall more than half of the boundaries were detected in at least two out of the three time points of infection. Similarly, when comparing

domain boundaries obtained from the infection data to the published boundaries (Dixon et al., 2012), around half of the boundaries between the two studies overlap. Even though, TADs have first been described five years ago and tremendous effort has been put in to develop algorithms to stably detect those domains and their boundaries, no gold standard exists yet. The performance of the existing algorithms greatly depends on the sequencing depth and the resolution used (Forcato et al., 2017). Ultimately, TAD boundaries possess high insulating properties and interaction frequencies across boundaries should be significantly reduced. This should result in clear separation into two “triangles” in cumulative heatmaps over TAD boundaries. Analysis of the average Hi-C interaction profiles around TAD boundaries, detected at individual time points or throughout the entire infection course, for all three stages of infection, revealed that most of the described TAD boundaries have strong insulation properties at all times of infection (Figure 4.7d), but were potentially just under the threshold for detection by HOMER.

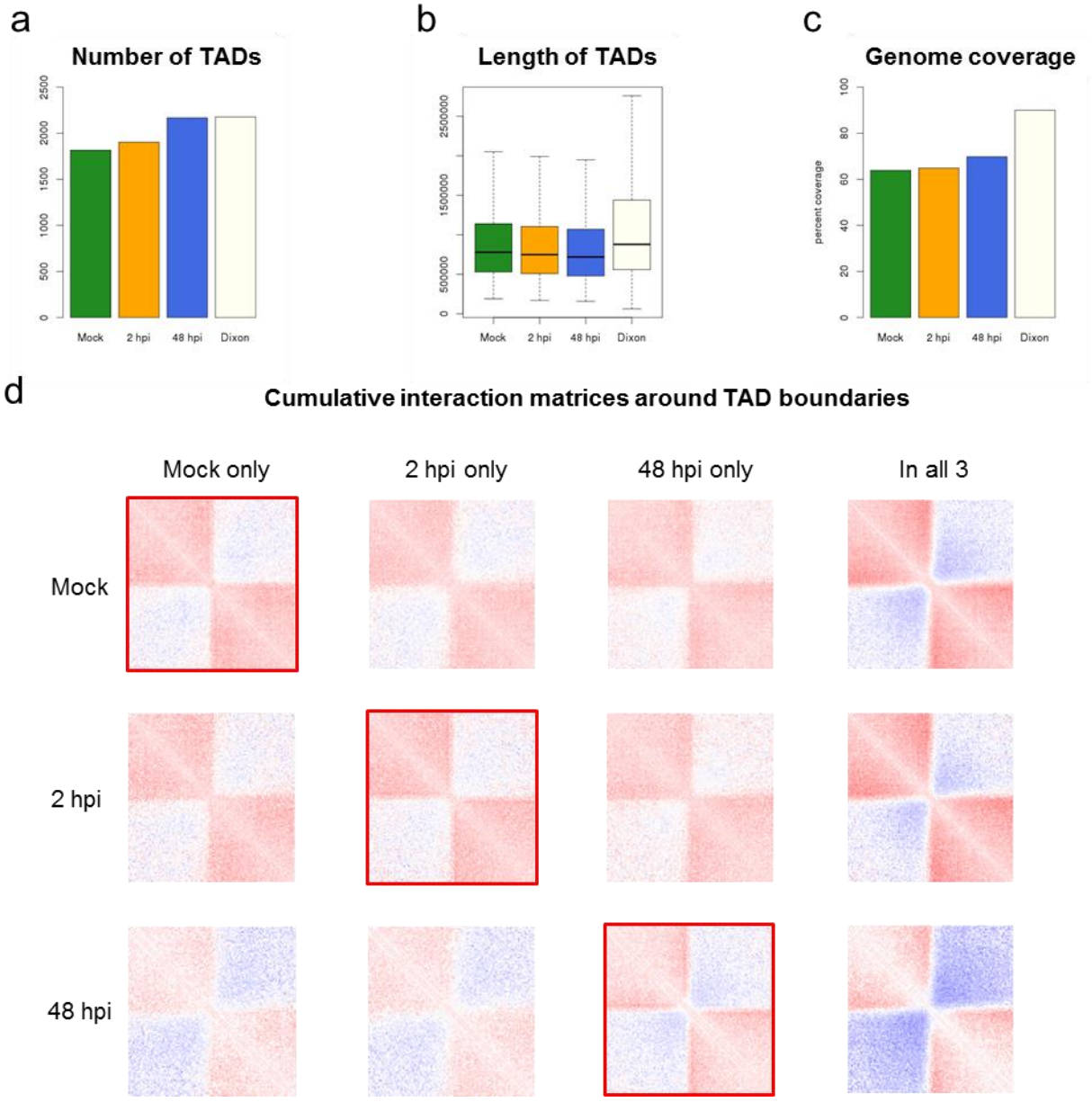


Figure 4.7 | TADs do not change boundary location upon lytic mCMV infection

TADs were calculated using the directionality index calculation (Dixon et al., 2012) for 40 kb bins with a distance cut-off of 2 Mb on combined replicates. **(a)** Barplot showing the observed number of domains at the different stages in infection and published numbers. **(b)** Boxplots showing the length distribution of the observed and previously described domains. **(c)** Bar chart displaying the resulting genomic coverage of TADs for the indicated samples. **(d)** Cumulative interaction profiles of 10 kb bins, in the region 1 Mb around TAD boundaries detected at different stages of infection (columns) in the different libraries (rows). Strong interactions are depicted in red, whereas blue represents weak interactions after normalisation and distance decay correction. Corresponding boundaries to the libraries they were detected in are boxed in red.

Thus, the union of TADs from all three stages of infection was formed to further investigate the phenomenon of striking differences of the strength of internal interactions within TADs. I calculated the cross-boundary ratio, which is the ratio of intra-TAD interactions over inter-TAD interactions (Figure 4.8a) for all individual TADs and monitored the differences throughout the

infection. No major changes could be detected between 2 hpi and the non-infected cells, but when comparing the ratios from NIH-3T3 with those calculated 48 hpi, a subset of domains clearly separating from the diagonal were detected (Figure 4.8b). TADs changing significantly in their cross-boundary ratio, under the null hypothesis that there are no changes between the two time points, were identified (Figure 4.8c). It had been shown that a loss of local interactions leads to a reduced physical compaction measured by FISH (Chandra et al., 2015). Therefore, it is reasonable to assume the inverse that an increase in local interactions i.e. within the same TAD, leads to physical compaction of that locus. Hence, the domains showing an increase in their cross-boundary ratio were denoted as closing TADs and the remaining TADs were annotated as non-responsive domains. To test whether specific properties of these closing or non-responsive TADs can be identified, I looked for an enrichment of DNA and chromatin features in both classes of domains. I found closing TADs to be depleted for genes, especially for expressed genes (as determined with a maximum of 0.7 RPKM cut-off across all time points; see chapter 3.3) and enriched for gene deserts, defined as 500 kb stretches of DNA without any annotated gene on them. Furthermore, the closing TADs seem to be enriched for the light A/T-rich isochores and increasingly depleted for the heavy G/C-rich H-isochores. Isochores are stretches of DNA of varying length with a more or less constant G/C-content (Costantini et al., 2009). To investigate the location of the closing TADS within the nucleus, I looked for the overlap with LADs determined in mouse embryonic fibroblasts (Peric-Hupkes et al., 2010) and found them to be located close to the nuclear periphery, interacting with the nuclear lamina. One finding, which is in good agreement with all of these findings above, is the overlap of closing TADs with late replicating domains (determined in MEFs (Hiratani et al., 2010)). This highlights that gene poor, late replicating, A/T-rich lamina associated domains are the domains that display more local interactions upon infection and are the ones being compacted. This corroborated the result, that LADs display a more compact chromatin structure than other non-lamina associated regions (Figure 4.6d). G/C-content, replication timing and the association with the nuclear lamina are not independent but also not entirely related. To further dissect the role of the nuclear lamina and other genomic feature, such as the G/C-content I separated the genome into 200 kb bins shifted by 50 kb, annotated them into LADs and inter-LADs and further subdivided the two categories based on their G/C-content (isochores). The proposed open chromatin index (OCI) provides an easy accessible measure for how local or distal certain genomic regions are (Chandra et al., 2015), by measuring the ratio of long-range and *trans* interaction to short range interactions (see 2.15.6). An increase in local interactions can only be observed for both of the A/T-rich lamina associated categories, while the rest of the genome shows a general trend towards more distal interactions. Being only A/T-rich or associated with the nuclear lamina alone is not sufficient for compaction upon lytic mCMV infection.

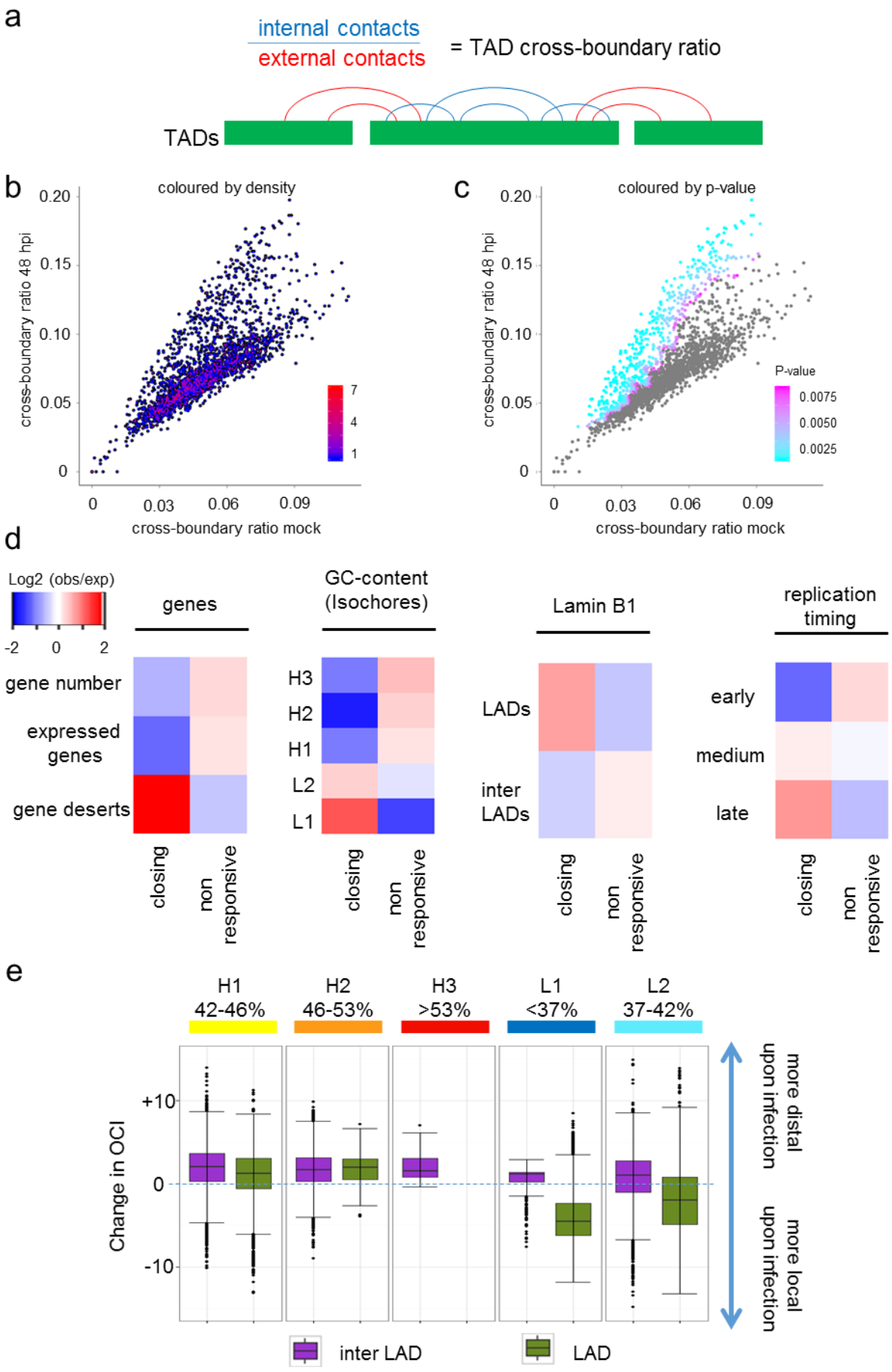

a
internal contacts
external contacts
= TAD cross-boundary ratio
TADs
b
coloured by density
cross-boundary ratio 48 hpi
cross-boundary ratio mock
c
coloured by p-value
P-value
0.0075
0.0050
0.0025
d
Log2 (obs/exp)
genes
GC-content (Isochores)
Lamin B1
replication timing
gene number
expressed genes
gene deserts
H3
H2
H1
L2
L1
LADs
inter LADs
early
medium
late
closing
non responsive
e
H1 42-46%
H2 46-53%
H3 >53%
L1 <37%
L2 37-42%
Change in OCI
more distal upon infection
more local upon infection
inter LAD
LAD

Figure 4.8 | Genomic properties predict the structural changes upon lytic mCMV infection

(a) Schematic illustrating the TAD cross-boundary ratio, which is used as a measure of compaction. **(b and c)** Scatterplot comparing TAD cross boundary ratio for the union of all TADs observed in the infection data between mock-infected and 48 hpi NIH-3T3, in **(b)** coloured by density and in **(c)** coloured by p-value. Colour scheme as indicated in the figure. This ratio increases for all TADs genome-wide at 48 hpi indicating genome-wide compaction. **(c)** A subset of TADs displays an even stronger increase than expected, compared to the rest of the TADs (depicted in different shades cyan). TADs were separated into closing and non-responsive TADs based on p-values (cut off: $p < 0.005$ for closing TADs and $p \geq 0.005$ for non-responsive TADs. **(d)** Enrichment for genomic features of closing and non-responsive TADs. Displayed is the log2 (observed /expected ratio), with red depicting enrichment over random and blue showing depletion compared to a random distribution. Closing TADs are enriched for inactive genomic features, such as gene and GC-poor, late replicating regions near the nuclear lamina. **(e)** Change in OCI for LADs and inter-LADs separated based on their GC-content (isochores). Only AT-rich LADs are being compacted upon infection and show more local interactions late in infection compare to non-infected cells.

In summary, these data highlight the importance and the conservation of TADs as structural units during infection. The observed compaction of regions near the nuclear periphery, described under 4.3.2.1 (Figure 4.6d), happens on the level of TADs and further seems to be dependent not only on the association with the nuclear lamina but also on the G/C content of the affected TADs.

4.3.2.3 A/B compartmentalisation of the host genome upon lytic mCMV infection

It is widely established, that at a Mb-scale, the genome is segmented into two, arbitrarily labelled A and B, compartments, in such a way that contacts within each set are enriched and contacts between sets are depleted, resulting in the observed plaid pattern of contact matrices (Figure 4.5b and e). It had been described that PCA can reveal to which compartment genomic regions belong (Lieberman-Aiden et al., 2009). The authors could further show that the A compartment correlates to active and more open chromatin, while the B compartment reflects more densely packed inactive chromatin. This was confirmed and extended by further studies, in which it was shown that changes in A/B compartmentalisation correlate with changes in transcriptional activity (Fortin & Hansen, 2015; Imakaev et al., 2012). In order to identify large genomic regions changing their transcriptional activity upon lytic mCMV infection, I calculated eigenvectors of 200 kb bins genome-wide. I observed chromosome specific patterns of the leading eigenvector, which were stable between replicates (genome-wide R^2=0.98). Furthermore, PCA was performed to compare genomic activity genome-wide for all 200 kb bins between the different stages of infection. No significant changes could be observed between non-infected cells and cells 2 hpi. At the late stage of infection, 48 hpi, substantial differences in either direction could be detected, as representatively shown for chromosome 10 (Figure 4.9a). More genomic regions are changing form the inactive B to the active A compartment genome-wide late in infection (Figure 4.9b). How to quantify best the differences in A/B compartments is still an open question. In previous studies, 0 has been used as the

threshold to differentiate between the two compartments (Lieberman-Aiden et al., 2009). It is not clear that functional differences exist exactly when the two eigenvectors have opposite signs, and I, therefore, chose a more conservative approach where values of the leading eigenvector had to change from lower than -5 to larger than 5, or vice versa. This approach ensures that the identified regions are indeed changing in activity. Using this highly conservative threshold, only eight 200 kb regions, containing only four genes, showed strong enough repression (Figure 4.9c). Only two of these genes possess annotated coding potential. *Gramb1d*, is involved in ectoderm formation and displays moderate upregulation late upon infection, whereas *LAMA4* (Laminin Subunit Alpha 4) is heavily down-regulated at the late stages in infection (Figure 4.9d). Laminins, a family of extracellular matrix glycoproteins, are the major non-collagenous constituent of the basement membranes. They have been implicated in a wide variety of biological processes including cell adhesion, differentiation, migration, signalling, neurite outgrowth and metastasis (Bonnans et al., 2014). In contrast to the low number of regions with decreased first PCA, exactly 100 of the 200 kb regions, covering in total 20 Mb of DNA and containing 287 genes, were found to increase their first PCA substantially (Figure 4.9c). The expression levels of these 287 genes significantly increased between non-infected cells and 48 hpi (Figure 4.9e; Fischer`s exact test; $p < 0.001$). No functional gene ontology term was found to be significantly enriched amongst those genes. To further dissect the genes identified, I separated them based on expression levels in non-infected cells into 45 highly expressed, 45 lowly expressed and 197 very lowly expressed genes and plotted their expression pattern over the time course of infection (Figure 4.9f). The highly expressed genes remained highly expressed and were associated with gene expression and translation. The lowly expressed genes significantly increased in expression between the non-infected cells and 48 hpi. Genes in this category were associated with metabolism. Notably, I could identify that developmental genes change from the B to the A compartment at late stages upon lytic mCMV infection. Although overall lowly expressed, these developmental genes displayed a significant increase in their transcriptional activity.

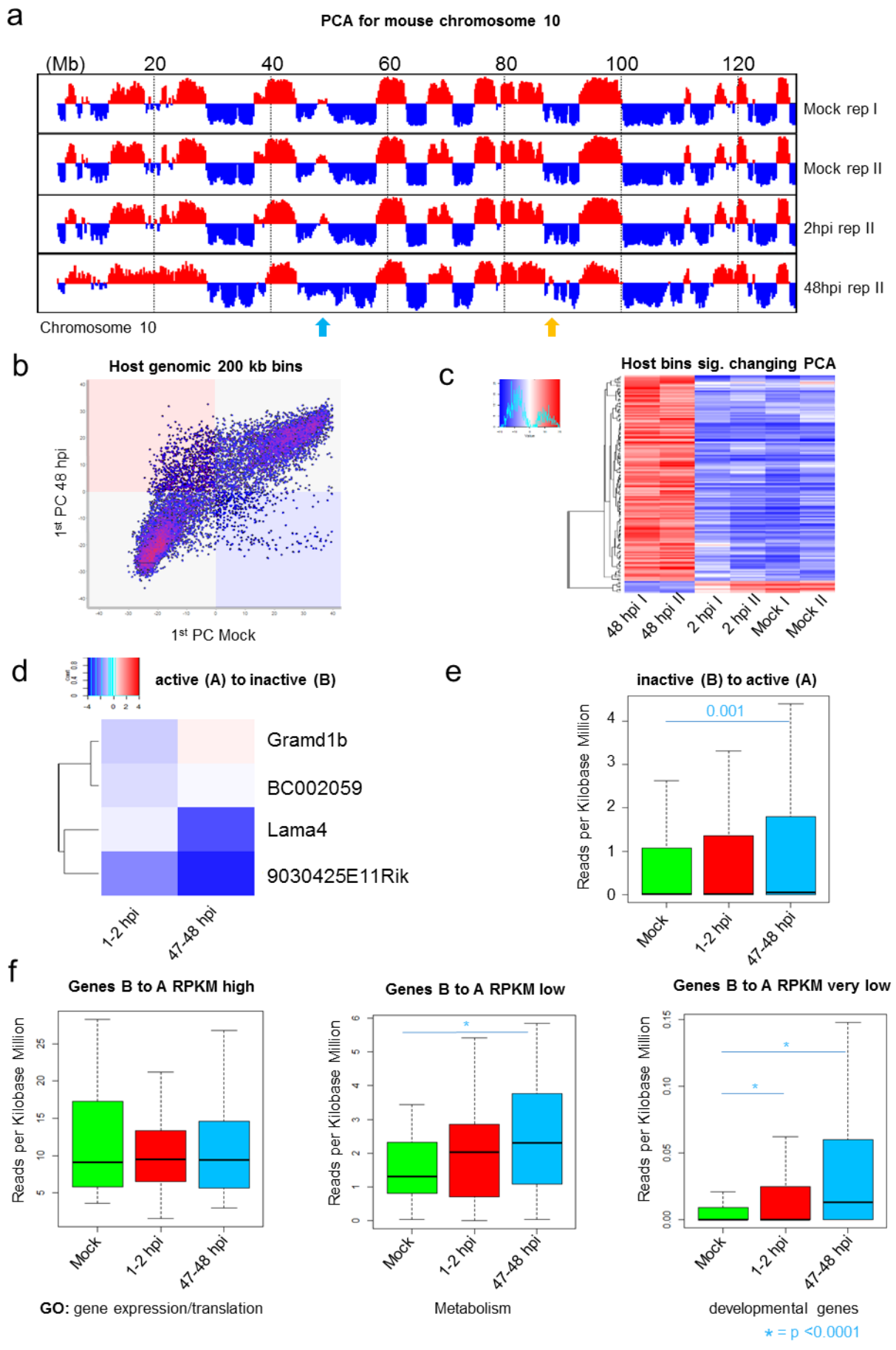

a
PCA for mouse chromosome 10
(Mb)
20
40
60
80
100
120
Mock rep I
Mock rep II
2hpi rep II
48hpi rep II
Chromosome 10
b
Host genomic 200 kb bins
1st PC 48 hpi
1st PC Mock
c
Host bins sig. changing PCA
48 hpi I
48 hpi II
2 hpi I
2 hpi II
Mock I
Mock II
d
active (A) to inactive (B)
Gramd1b
BC002059
Lama4
9030425E11Rik
1-2 hpi
47-48 hpi
e
inactive (B) to active (A)
0.001
Reads per Kilobase Million
Mock
1-2 hpi
47-48 hpi
f
Genes B to A RPKM high
Genes B to A RPKM low
Genes B to A RPKM very low
Reads per Kilobase Million
Mock
1-2 hpi
47-48 hpi
GO: gene expression/translation
Metabolism
developmental genes
* = p <0.0001

Figure 4.9 | Specific loci switch between open and closed A/B compartments only late in infection

(a) 1st principal component of Hi-C data for 200 kb bins displayed for chromosome 10 in the two biological replicates for non-infected cells and for the second replicate for 2 hpi and 48 hpi. PCA is very stable between replicates and changes in either direction can only be observed at late stage of infection. Exemplary changes from A to B compartment (inactivation) are indicated by the orange arrow, whereas exemplary changes from B to A compartment are depicted by a light blue arrow. **(b)** Scatterplot comparing genome-wide changes in 1st principal component for all 250 kb bins reveals changes in either directions can be observed genome-wide. **(c)** ANOVA revealed that changing form B to A compartment (activation) late in infection is a more common feature of lytic mCMV infection, compared to inactivation. **(d)** Heatmap of log2 (fold changes) in gene expression compared to non-infected expression levels for genes overlapping regions changing from A to B compartment indicates down-regulation of those genes. **(e)** Bar chart indicating that genes contained within loci changing from B to A compartment are significantly up-regulated, but show wide range of expression levels. **(f)** Genes overlapping regions becoming active were further subdivided based on their expression levels in non-infected cells. Initially highly expressed genes in those loci stay highly expressed and are associated with the gene ontology terms: gene expression and translation, whereas lowly expressed genes in NIH-3T3 are significantly upregulated late in infection and relate to metabolic functions. Notably, among the non-expressed or very lowly expressed group are genes that are significantly upregulated and that are enriched for developmental genes.

Taken together, these data suggest that, despite the global compaction of host chromatin, more regions are changing from the inactive B compartment to the active A compartment, including genes important for embryonal development. The dramatically down-regulated gene *LAMA4*, is changing from the A to the B compartment, which corroborates the functional relation between genome organisation, observed by compartmentalisation, and genome function, detected by transcriptional activity.

4.3.3 SCRiBL allows to assess individual promoter enhancer loops

So far, I have explored how the host cellular genome organisation changes upon lytic mCMV infection on a more global architectural level and reported compaction of large domains near the nuclear periphery. Mammalian genomes harbour a large number of *cis*-regulatory elements that have been shown to be instrumental for gene expression during development and other cellular processes (Bulger et al., 2011). To determine regulatory elements for profoundly regulated genes upon viral infection and to further study the dynamic changes of promoter-enhancer interactions upon lytic mCMV infection, I employed sequence capture of regions interacting with bait loci (SCRiBL) in biological replicates of non-infected NIH-3T3, and virus infected cells at stage 2 hpi and 48 hpi. The key steps in the protocol are outlined in Figure 4.1. I captured genomic regions surrounding genes that showed induction of gene expression 1-2 hpi followed by strong viral counter-regulation (three NF-κB regulated genes and three IF-regulated genes), surrounding seven constant up- and surrounding six continuous down-regulation genes. Five control regions including histone genes and developmental genes that did not display a change in compartmentalisation were included in this assay. The BACs used for bait generation in this study are listed in Table 2.4. Compared to my pre-capture Hi-C

libraries this sequence capture step enriched more than 200-fold for host genomic sequences and more than a thousand fold for the viral genome. Furthermore, the quality of the sequenced libraries increases for host genomic regions by including this capture steps (Figure 4.4), suggesting that it is enriching for valid di-tags rather than non-informative ligation events. Capturing specific regions markedly reduces the overall complexity of the pre-capture Hi-C libraries, meaning at comparable sequencing depth, the resolution is much higher for the targeted loci (Figure 4.10a). In order to obtain an equivalent number of reads falling onto promoters of interest, a Hi-C library would need to be sequenced up to 200-fold greater depth (Figure 4.10b). The theoretical maximum enrichment is dependent on the size of the captured region. For example, using the 5.6 Mb SCRiBL capture system described here, the maximum theoretical enrichment is~ 580-fold over Hi-C. While capturing the viral genome, with its ~230 kb, alone can theoretically achieve a 12,000-fold enrichment. To demonstrate that SCRiBL enables the analysis of long distance interaction with non-capture regions, I obtained virtual 4C profiles form the restriction fragment containing the *Hist1h4d* TSS at all three stages of infection (Figure 4.10c). A clear local distance decay within the captured regions was found at all three time points. Further the *Hist1* histone cluster on chromosome 13 displays a long-range interaction spanning over more than 1 Mb and even over another captured region, containing a vomeronasal receptor cluster. This highlights that SCRiBL libraries are not dominated by the captured sequences and rather reflect genomic preferential interaction. With this increased power, SCRiBL enables analysis of captured regions, including promoters of interest at restriction fragment level.

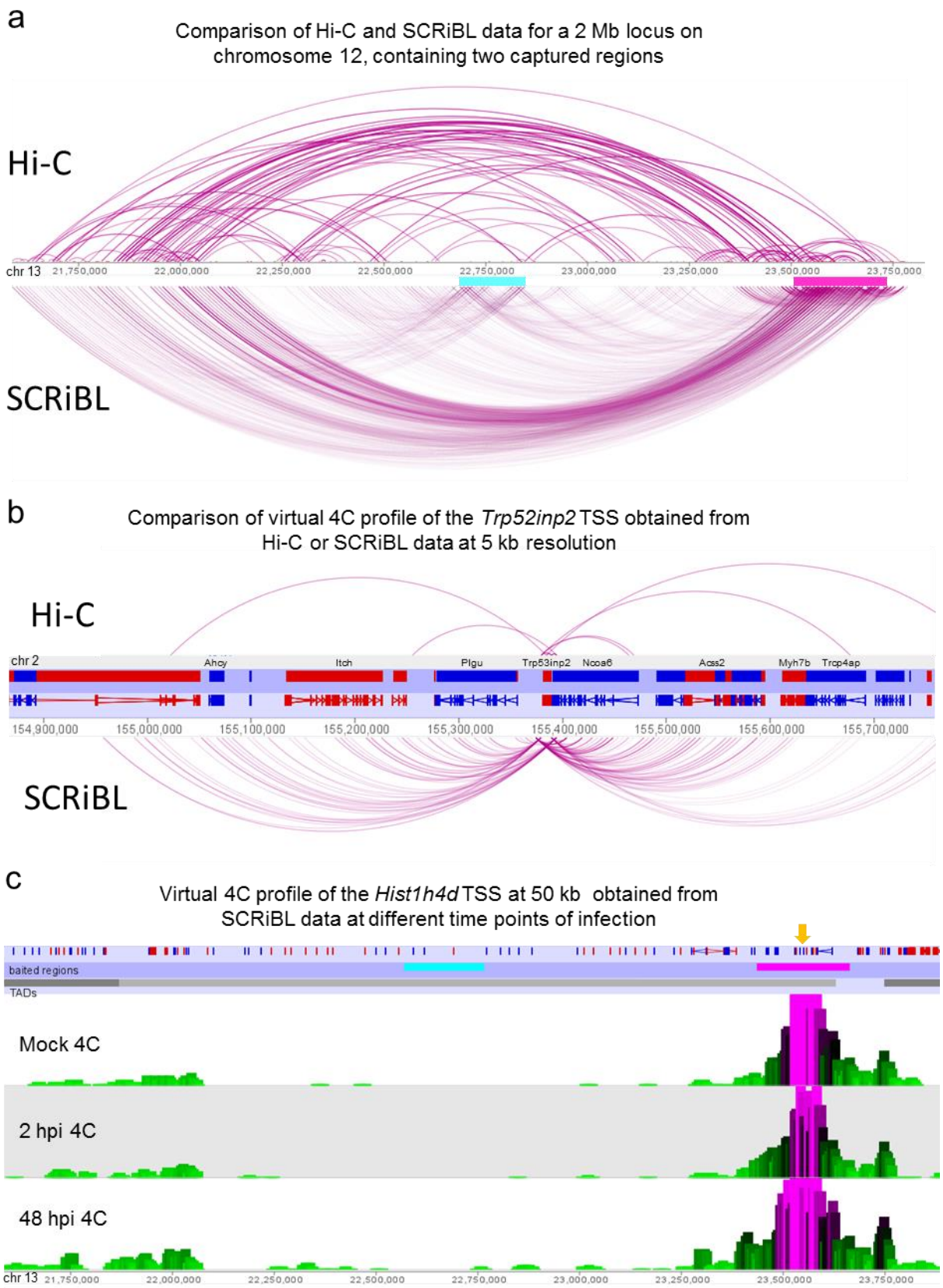


Figure 4.10 | SCRiBL enriches Hi-C libraries specifically and significantly for regions of interest

(a) The chromosomal interactome of the *Hist1* locus in NIH-3T3 at 5 kb resolution. Shown are unfiltered read pairs from Hi-C data (*top*) for a 2 Mb region containing the *Hist1* captured locus (magenta rectangle) on the far right and

read pairs from SCRiBL (*lower*). The captured region containing the vomeronasal receptor cluster is indicated by the cyan rectangle. Hi-C and SCRiBL data sets were adjusted to the same number of overall sequence reads. Interactions are displayed using the WashU EpiGenome Browser. **(b)** Virtual 4C profiles at 5 kb were obtained from the TSS of *Trp53inp2* in Hi-C and SCRiBL datasets and displayed in the WashU EpiGenome Browser. **(c)** Virtual 4C profiles were obtained from the histone *Hist1h4d* TSS (orange arrow) in the SCRiBL data from the different time points of infection and reads of the interacting ends were binned into 50 kb bins. Clear interaction within the captured region (magenta rectangle) can be seen with a strong local distance decay. The previously described, cell-type invariant interaction between the *Hist1* histone cluster 1.9 Mb away was detected, whereas a captured region in between (cyan rectangle) does not interact with any of the two.

Therefore virtual 4C profiles at 10 kb resolution for all 24 TSS of interest (see Table 2.4) were generated for the three different time points of infection using Seqmonk. I could identify several short and long range interactions, within and outside of the captured regions for all 24 genes (supplementary figure 4). Surprisingly, I found that virtually all established contacts in non-infected NIH-3T3 between the TSS of interest and all other 10 kb regions in the host genome remain largely unchanged, with correlation scores between time points of the same biological replicate being comparable to correlation scores between replicates (Figure 4.11a). The comparison of correlation scores between time points of the same replicate within the four different classes of expression kinetic (control, up, down and NF-κB induced) reveals that there are no major differences in the v4C profiles (Figure 4.11b). Most of the interactions between promoters and their potential regulatory elements are being pre-established and are less dynamic than anticipated. This is especially surprising for the highly induced genes where a boost in gene expression through a newly formed interaction with an enhancer was a plausible underlying mechanism. To illustrate one of those pre-existing contacts, the 4C profiles obtained from the TSS of the *Nfkbia* gene, which is regulated by the TF NF-κB and shows strong induction 1-2 hpi followed by counter-regulation by viral gene products, are depicted in Figure 4.11c. Several prominent interacting regions could be identified and are highlighted in the Figure 4.11c by coloured rectangles. The black box immediately upstream of the TSS is by definition close by in the 3D space, shows high accessibility and most likely represents the PPR. More intriguing is the interacting region ~40 kb downstream of the gene (depicted by the green rectangle in Figure 4.11c), which shows accessibility throughout the entire infection, although dramatically reduced at the late stage of infection. Further, in non-infected cells and 2 hpi transcription can be observed at this interacting regions, which again is dramatically reduced 48 hpi. Active enhancers produce functional non-coding RNAs (eRNAs), which are instrumental to the enhancers function, at least in the case of p53 mediated transcription (Melo et al., 2013). Investigation of previously published results (Dixon et al., 2012) suggested, that both interacting regions downstream of the gene are carrying active enhancer marks, such as H3K4me1 combined with H3K27Ac in non-infected primary MEFs (data not shown).

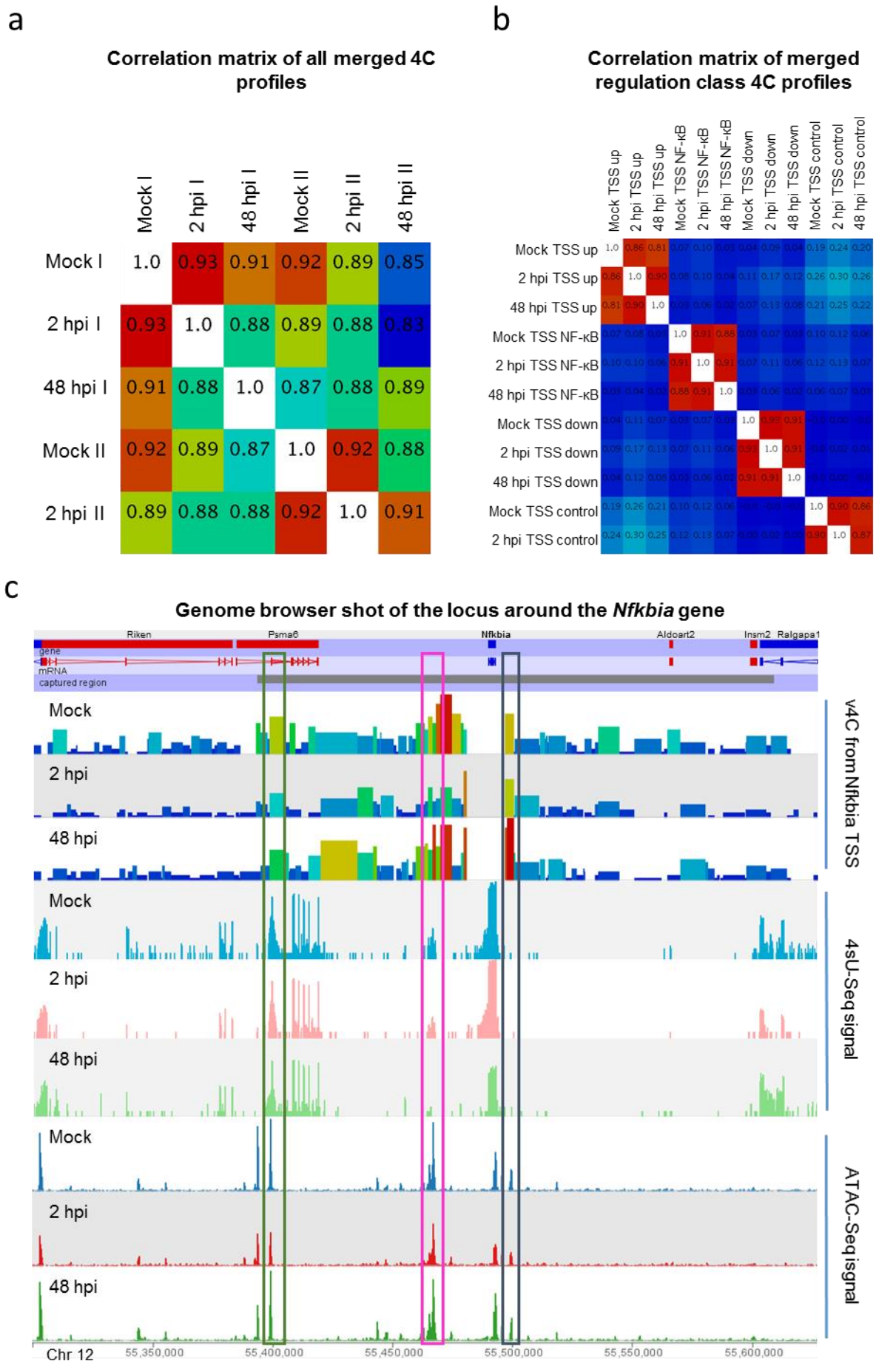

a
Correlation matrix of all merged 4C profiles
Mock I
2 hpi I
48 hpi I
Mock II
2 hpi II
48 hpi II
b
Correlation matrix of merged regulation class 4C profiles
Mock TSS up
2 hpi TSS up
48 hpi TSS up
Mock TSS NF-κB
2 hpi TSS NF-κB
48 hpi TSS NF-κB
Mock TSS down
2 hpi TSS down
48 hpi TSS down
Mock TSS control
2 hpi TSS control
48 hpi TSS control
c
Genome browser shot of the locus around the Nfkbia gene
Riken
Psma6
Nfkbia
Aldoart2
Insm2
Ralgapa1
gene
mRNA
captured region
Mock
2 hpi
48 hpi
v4C from Nfkbia TSS
4sU-Seq signal
ATAC-Seq signal
Chr 12
55,350,000
55,400,000
55,450,000
55,500,000
55,550,000
55,600,000

Figure 4.11 | Pre-existing loops are a common phenomenon and exert their function through changes in enhancer activity

(a) Matrix depicting the Pearson`s Correlation between merged virtual 4C (v4C) profiles obtained from all TSS of interest (as defined in Table 2.4) at 10 kb resolution, for both biological replicates. **(b)** Pearson`s Correlation matrix depicting the correlation between v4C profiles obtained from biological replicate 2 for merged expression categories. **(c)** Genome browser shot of the 350 kb region around the *Nfkbia* gene on chromosome 12 showing virtual 4C data with BglII restriction fragment resolution obtained from the *Nfkbia* TSS fragment, 4sU-Seq and ATAC-Seq data for non-infected NIH-3T3, cells 2 hpi and 48 hpi from merged replicates. All data were normalised to sequencing depth, thus enabling direct comparison between the three time points within a given data set. Three pre-existing contacts (coloured rectangles) can be observed in NIH-3T3, which persist throughout the infection. The contact 40 kb downstream of the *Nfkbia* gene (magenta rectangle) displays striking behaviour, by showing dramatically reduced non-coding transcription and accessibility late in infection, which correlates well with expression levels of the gens itself.

Taken together, SCRiBL enables the interrogation of locus specific genomic architecture and looping. In the case of lytic mCMV infection, it could illustrate that a pre-wired promoter enhancer contact between a highly regulated gene and at least one distal regulatory element ~40 kb away exists. Integrative analysis of accessibility data, information of nascent transcription and structural data suggests that this pre-existing and maintained loop is able to exert its function by altering its activity.

4.3.4 Hi-C contains spatial information of the virus

Next, I addressed how the viral genome itself is organized and if I can detect any specific interactions with the host genome at any stage of the infection. To detect host-pathogen DNA:DNA interactions, for the early 2 hpi time point, the SCRiBL dataset was analyzed, whereas for the late stage of infection Hi-C was sufficient, as half of the reads fall on the viral genome already without capture. Using GOTHiC (Genome Organisation Through Hi-C) (Mifsud et al., 2017) several host genomic 250 kb bins were found to significantly interact with the entire viral genome at different stages of infection. Hundred thirty-nine genes were found to only interact with the virus DNA early in infection and 1355 genes interacted with the viral genome late in infection, whereas 294 genes were found to contact the viral genome throughout the infection (Figure 4.12a). Concomitant with the increase in viral copy number, an increase in contacted host genes could be detected 48 hpi. None of the three categories showed enrichment for specific cellular functions determined by gene ontology terms. Strikingly, at all given time points during infection the viral genome is not in close proximity to genomic regions near the nuclear lamina, at the nuclear periphery (Figure 4.12b). To investigate further, if the genomic interactions between the host and the virus are mediated by specific loci on the viral genome, I calculated the *trans*/*cis* ratio for individual restriction fragments on the viral genome. All interaction between the virus and the host are by definition interchromosomal, as the virus does not integrate but resides as an extrachromosomal

episome. I observed that the viral genome appears to be unstructured without exhibiting distinct *cis*/*trans*-rich regions at early time points of infection. At 2 hpi the viral genome forms more interactions with the host than 48 hpi; 95 % of all viral interactions are with the host 2 hpi, whereas 48 hpi only 50 % of the viral interactions are directed towards host cellular DNA. This is consistent with the formation of VRCs during lytic infection. I had speculated that the incoming viral genomes might preferentially localize to transcription factories containing NF-κB or IF-regulated genes as the mCMV immediate early promoter contains binding sites for the associated TFs e.g. the CMV MIEP has four NF-κB-binding sites and activation of NF-κB is critical for MIEP activation and eventual expression of all viral genes (Speir et al., 1998). Despite the relatively large number of viral genomes entering the cell, most of which may actually not be transcribed at all, this should result in distinct trans-rich regions. In contrast, we observed that the interaction of the viral genomes at both 2 hpi with the host chromatin resembled a random ligation pattern rather than distinct interactions. In contrast, a well-defined *cis*/*trans* interaction pattern was observed at 48 hpi. Considering that by 48 hpi virus replication has resulted in a >1,000 fold genome replication resulting in a huge amount of viral genomes within the nucleus, it is surprising to find the majority of all trans interaction between the virus and the host are formed by the 7.2 kb viral intron. This intron shows high expression rates at the late stages of infection.

Next, I aimed to identify the structure of the viral genome at the different stages of infection and to identify potential cis-regulatory elements within the viral sequence. Therefore, distance and sequencing depth corrected interaction matrices of 5 kb bins of the viral genome were generated and analysed.

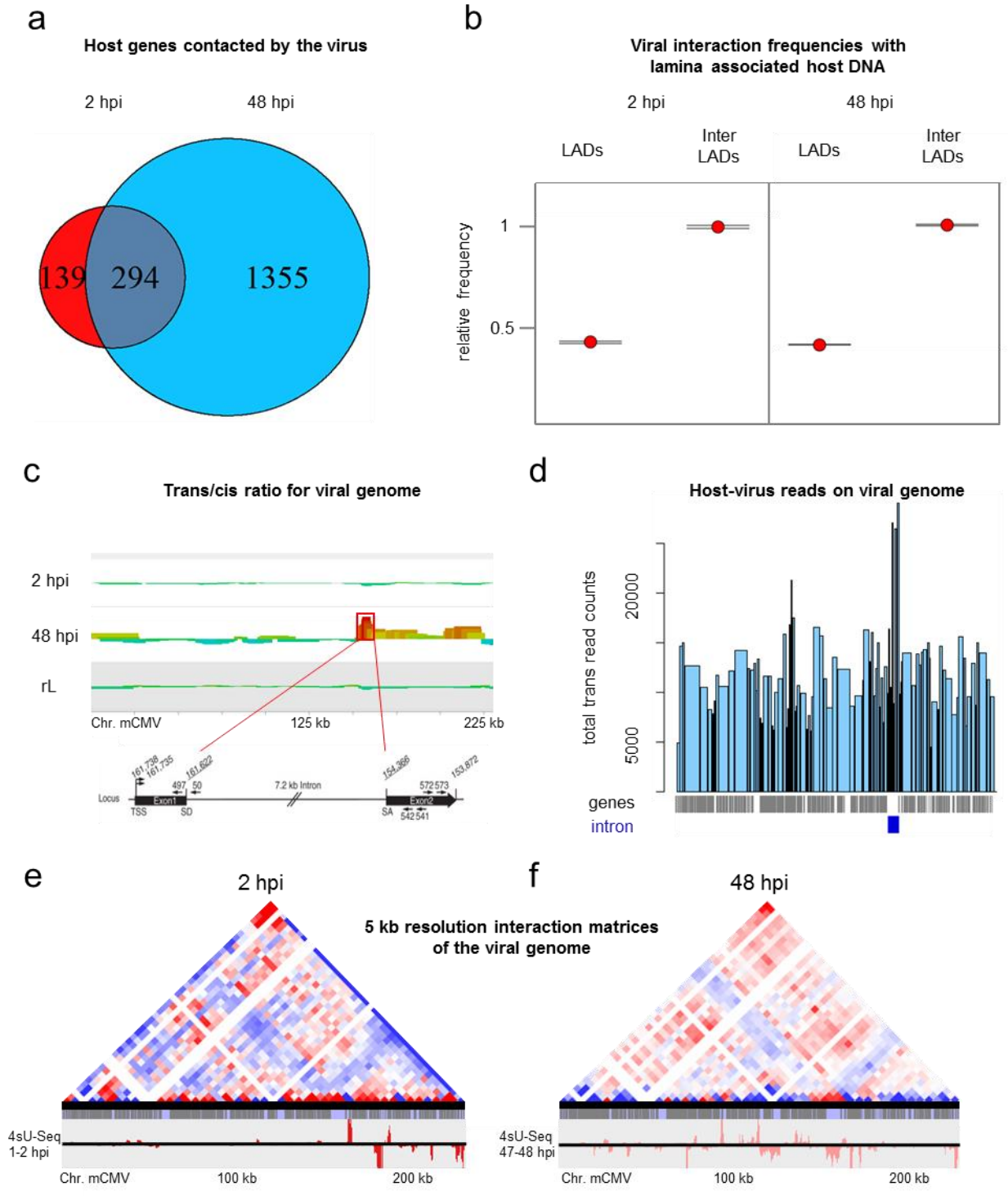


Figure 4.12 | Interaction profiles of the viral genome and between the virus and the host DNA

Virtual 4C profiles from the entire viral genome were obtained 2 hpi from SCRiBL experiments and 48 hpi from Hi-C data (both combined replicates) and binned in 250 kb regions. Significantly interaction bins were identified using GOTHiC. **(a)** Venn diagram displaying the number of genes within significantly interacting bins at the two time points of infection. No enrichment for specific functional annotations (GO terms) or TFBS could be identified. **(b)** Star wars plot showing the normalised frequencies of finding the interacting host genomic locus to be a LAD or an inter-LAD. Clear enrichment for contacting inter-LADs at both stages of infection is visible. **(c)** Bins of 5 neighbouring BglII fragments on the viral genome were merged, subjected to trans/cis quantitation and visualised in Seqmonk. No

outstanding region of clear interchromosomal interacting region could be identified, although the overall *trans* ratio of the viral genome was 95 % at that time of infection. At the late stage of infection, an obvious peak of this ration can be detected, directly matching the 7.2 kb stretch encoding for a viral intron. **(d)** This is not an artefact of the measurement used, but actually reflects that many interchromosomal reads are stacking up over this region, indicating that this region is mainly contacting the host genome late in infection. Interchromosomal contact matrices of the viral genome at 5 kb resolution were obtained at **(e)** 2 hpi from SCRiBL data and **(f)** 48 hpi from Hi-C data (both combined replicates).

This revealed that at both time points reported, both terminal ends of the viral genome, as displayed when the viral genome is shown as a linear genomic stretch, strongly interact with each other (Figure 4.12e and f). This is in agreement with the circularization of the viral genome immediately upon entering the nucleus, later on enabling the rolling cycle DNA replication leading to head to tail concatemers. Furthermore, clear depletion of the large viral intron for interactions in *cis*, at both time points of infection, was observed. At the early stage of infection, broader domains of the viral genome are in close spatial proximity, which seem to be the transcribed loci at that time during infection. At 48 hpi all of the viral genome seems to be in close spatial proximity to each other. Strikingly, the usually observed diagonal, indicating strong close cis interactions appears to be nonexistent.

In summary, Hi-C and SCRiBL chromosomal interaction plots revealed that initially the entire viral genome is in close proximity to open host chromatin, whereas at later stages of infection most of the interactions are mediated by a noncoding locus of the viral genome, which further seems to be depleted of interactions in *cis*.

4.4 Discussion

Due to the profound changes seen within DNA-stained lytically mCMV infected cells, a number of studies have attempted to capture the key drivers of mCMV infection using epigenomic and microscopic techniques. Using Hi-C, I have been able to generate a comprehensive description of spatial changes in genome architecture within the infected nuclei on a global scale. SCRiBL highlighted the existence of pre-established regulatory loops between putative enhancers and promoters and integration of other epigenomic data revealed that host regulatory elements potentially exhibit their function during infection via changing their activity.

In this thesis, four Hi-C libraries were generated from different time points during lytic mCMV infection of murine NIH-3T3 fibroblasts to study the three-dimensional changes in nuclear architecture as well as the nuclear localization and organization of mCMV genomes. The early time point, 4 hpi, using a low MOI of 0.5 infectious particles per cell resulted in interaction profiles resembling the ones obtained from non-infected NIH-3T3 and therefore was excluded from further analysis. High-throughput experiments measuring the average of a population need a homogeneous cell population to start with; otherwise, minor effects will be diluted too

much to be detectable. Furthermore, the three informative Hi-C libraries were enriched using biotinylated RNA baits for 24 host genomic loci of 150-250 kb each, covering roughly 5 Mb in sum, and the viral genome. This enabled a >200-fold enrichment in sequencing depths over conventional Hi-C, thus providing an ideal tool to study changes in nuclear architecture at high resolution.

Genome-wide heatmaps and intrachromosomal interaction matrices revealed the expected diagonal pattern, corresponding to strong and abundant *cis*-interactions on all chromosomes. These heatmaps further indicated that NIH-3T3 cells harbor a t(4:16)-translocation, visible as a distinct spot in the heatmaps, which is in agreement with the lately published karyotype of NIH-3T3 (Leibiger et al., 2013). This study reported five derivative chromosomes consisting of two different chromosomes each. Heatmaps obtained from all time points indicate that there are additional genomic aberrations. NIH-3T3 have been passaged for multiple decades and due to immortalization, display an instable karyotype and Hi-C can be employed to detect chromosome rearrangements and copy number changes (Harewood et al., 2017). All of the observed genomic instability was unchanged throughout the experiments described here, ruling out the possibility that they can explain the observations made. Furthermore, I am comparing changes between the different time points rather than making statements about specific significant interactions, thus the anomalous karyotype is not likely to affect the results presented. The same is true for the reported copy number deviations.

Moreover, visual inspection of contact matrices suggested that the overall folding of the host genome is largely preserved throughout infection. Considering the dramatic changes in nuclear architecture visible by light microscopy during lytic mCMV infection, this was quite surprising. Nevertheless and consistent with these results we observed that the similarity to the mock-infected cells decreased throughout infection. This is not due to a dramatic structural change but rather appears to be due to an increase in *cis*-interactions and, therefore, due to a stronger regulation and tighter packing of the genome, especially on the level of specific TADs. I consistently find an A/T-rich compartment of the genome gaining internal contacts upon infection. This compartment is enriched in L1 and L2 isochores and overlaps compacted regions near the nuclear periphery, which are interacting with the nuclear lamina. This correlation between compaction, isochores and the nuclear lamina suggests that isochores may have physical properties beyond simple sequence recognition that allow the genome to rearrange its architecture during a stress response. A relationship between isochore structure and stress response would have important implications for evolution and could go some way to explaining the differences between the integration and fixation rates seen for some repetitive elements (Costantini et al., 2009). Additionally, Chandra et al 2015 have reported that A/T-rich DNA regions associated with the nuclear lamina are experiencing the most dramatic structural

changes during senescence by losing compaction. The authors reported the relocation of these genomic regions from the nuclear periphery towards the centre, where they associate into senescence-associated heterochromatic foci (SAHFs), indicating a specific structural role of A/T-rich LADs (Chandra et al., 2015). Thus, it would be interesting to characterize these A/T-rich LADs further, especially to see whether they are a mCMV infection-specific feature or whether they exist in other cellular states and stress responses.

Despite the compaction of L-LADs, the rest of the genome seems to become more distal interacting upon infection, which is rather surprising considering the massive reduction of the inter-chromatin space and the resulting compaction of the genome observed by microscopy (Gibbs et al., 2013). Furthermore, PCA revealed large regions of the genome changing in A/B compartmentalisation. Surprisingly, considering the microscopy data, but consistent with transcriptional data presented in chapter 3, only very few regions changed from the active A to the inactive B compartment. These were gene poor and only contained four genes, among the laminin alpha 4 gene *LAMA4*, which, in agreement with the published concept of A/B compartments (Lieberman-Aiden et al., 2009), showed a strong reduction in expression at 48 hpi and is part of the extracellular matrix. Interestingly, hCMV infection can be linked to cardiovascular diseases and has been shown to down-regulate extracellular matrix (ECM) proteins in both human fibroblast and smooth muscle cells (Reinhardt et al., 2006). The ECM has traditionally been regarded as an inert scaffold that merely provides mechanical support, but it has become increasingly evident that it also regulates growth, death, adhesion, migration, invasion, gene expression and differentiation of cells (Bonnans et al., 2014). Therefore, modulating the ECM may provide a mechanism for detachment and spreading of infected cells through the blood circulation, thus leading to cell bound virus spread.

Intriguingly, I found many regions changing from the inactive B to the active A compartment, comprising 293 genes, that displayed varying expression levels but as a group showed significant induction of gene expression. Further dissection into expression categories revealed that especially genes related to metabolism are upregulated. Viral DNA replication and particle production is intense on the cellular resources, therefore an enhanced metabolism is desirable for the virus to facilitating its needs (Vastag et al., 2011). Fascinatingly, non-expressed developmental genes were found to be located in genomic regions changing from B to A compartment. Congenital CMV infection is known to cause developmental aberrations in the central nerve system (CNS), leading to mental retardation and visual and hearing impairments. So far mechanisms including apoptosis, changes in cell cycle and neuroinflammatory processes have been proposed, but neuropathogenesis of hCMV infection is not fully understood (Cheeran et al., 2009). The data presented in this chapter would point towards a deregulation of developmentally important genes by genomic architectural changes

as an additional mechanism. The observed transcriptional changes in fibroblast were only minor, but in a more susceptible cell type, it is tempting to speculate that the changes could be more dramatic or at least minor changes can have detrimental effects, as these developmentally important genes are tightly regulated.

Since previous studies have shown that gene activation by enhancers is accompanied by alteration of chromatin interaction (Schoenfelder et al., 2010a; Smallwood & Ren, 2013), I expected that at shorter distance, virus-induced binding of NF-κB to enhancers would induce looping interactions that bring the distal enhancers into proximity with target genes. I therefore employed a novel, to date unpublished, technique called SCRiBL. The concept of this technique was originally inspired by an exon sequence capture method (Gnirke et al., 2009). Recently, multiple protocols including sequence enrichment steps have been established (Dryden et al., 2014; Hughes et al., 2014; Mifsud et al., 2015; Schoenfelder et al., 2015a). One caveat of all of these is that only a subset of all interactions in a 3C or Hi-C library can be enriched and one has to choose the regions of interest very carefully in order not to miss any interesting interactions. Additionally, biotinylated custom-made RNA baits are required, which are commercially available, but are still cost intensive. Here, I utilize a protocol that circumvents the necessity of purchasing those RNA species by generating them in house. The downside of this approach is that it is hard to determine statistically significant interactions, as no statistical model nor analytical tool exists. Nevertheless, comparison between different libraries enriched with the same capture system is possible. Furthermore, it seems that including the sequence capture step is not only increasing the sequencing depth by reducing the complexity of Hi-C libraries, but also increases the quality of the captured libraries by capturing preferentially valid, *cis*-interacting di-tags (Figure 4.4b). Most invalid sequences in my Hi-C libraries arise from the same internal fragment, which is genomic non-ligated DNA. By enriching for restriction fragment ends, as done in SCRiBL, these DNA fragments will get depleted from the captured library. When comparing SCRiBL to more traditional 3C based, biotinylated RNA-free techniques several advantages are obvious. 5C generates high-resolution chromosomal interaction landscape maps of megabase-size genomic regions but is not capable of capturing interactions involving DNA sequences outside the 5C target region(s). ChIA-PET combines antibody-mediated precipitation with ligation to map chromosomal associations. This depends on the availability and efficiency of suitable affinity reagents and restricts bait choice to genomic regions occupied by a protein of interest. SCRiBL on the other hand is relatively unbiased, enabling the comparison of chromosomal interaction profiles for genomic regions regardless of the cell type or differences in protein occupancy. A similar approach, but purchasing custom made biotinylated RNA oligonucleotides to enrich 3C libraries for specific genomic interactions, termed Capture-C (Hughes et al., 2014), has been

described. When comparing Capture-C and SCRiBL, I found that the percentage of sequence reads representing genuine chromosomal interactions (valid read-pairs, determined by HiCUP, data not shown) is about 10-fold higher in Capture Hi-C compared to Capture-C, presumably due to the fact that genuine ligation junctions are not pre-enriched in 3C libraries.

The high-resolution interaction profiles for specific loci, generated by using this method, suggest that in general, enhancer-promoter interactions already form in untreated cells; and these pre-existing DNA-structures are not significantly altered by transient activation or repression of enhancers. Recently, pre-existing promoter-enhancer looping was reported at several loci induced by p53, FOXO3 and glucocorticoid receptor using 4C approaches (Melo et al., 2013; Phillips-Cremins et al., 2013; Tan et al., 2012). Lately, genome-wide contact maps of TNF-alpha treated cells were made available (Jin et al., 2013). This revealed that this trend is a commonly observed phenomenon and further even predicts the level of gene induction. Here, my results further show that a target of an enhancer is already hardwired into the chromatin architecture, prior to the stimulus. This corroborates the proposed model, in which controlling the accessibility of enhancers at pre-established enhancer-promoter contacts adds an additional layer of regulation (Jin et al., 2013), potentially enabling fast and robust regulation. A study employing RELA ChIA PET, which is a subunit of the predominantly found NF-κB isoform, could identify that in human endothelial cells about half of the regulated genes upon TNF stimulation are already in close proximity to their enhancers prior to the treatment. Furthermore, they found that up- and down-regulated genes engage in preformed spatial interactions that bind RELA at 30 min after treatment (Kolovos et al., 2016). This diverse regulatory output is in line with a role for NF-κB as both an activator and a repressor and with the formation of NF-κB–driven spatial networks (Kuznetsova et al., 2015; Papantonis et al., 2012). It would be interesting to see if NF-κB binds at those distal promoter interacting regions upon mCMV infection and whether it remains bound at those sides. Furthermore, if NF-κB also exerts this bimodal role upon mCMV infection, or whether other TFs are being recruited and take over the function is not known. Since there are no specific candidates yet, an unbiased screen for potential TFs bound is required. Deeply sequenced ATAC-Seq libraries have been proposed to possess the potential to reveal TF footprints (Buenrostro et al., 2013), but the currently feasible sequencing depth is still far from the saturating level required for the detection of most TF footprints by ATAC-Seq (Sung et al., 2016).

Hi-C and SCRiBL libraries described here not only contain the 3-D information of the host genome but also structural information of the viral genome itself and the spatial relation between the two. In this thesis, I provided interchromosomal interaction maps of the viral genome at 5 kb resolution. These are, to the best of my knowledge, the first interaction heatmaps of a large DNA virus, documenting spatiotemporal changes with the ongoing

infection. Herpes viral genomes are densely packed with genes, which can overlap and reside on both strands. Hence, 5 kb resolution is not sufficient to provide a detailed picture on promoter contacts to potential regulatory regions, nor on specific gene networks contacts. Nevertheless, larger regions, which are in close 3D proximity to each other, were reported. These regions were reported to be expressed, thus suggesting the existence of a viral transcription compartment. Transcriptionally active genes in mammals have been proposed to associate with transcription factories, discrete nuclear sites of nascent RNA production and concentrated transcriptional components, such as RNA polymerase (Iborra et al., 1996; Osborne et al., 2004; Osborne et al., 2007; Schoenfelder et al., 2010b). Furthermore, I was able to document the interaction between the two ends of the viral genome, at both time points. At early stages of infection, the viral genome circularizes, enabling the rolling-cycle replication of the viral DNA later on, which produces head-to-tail concatemers that are subsequently cleaved into monomeric units and packaged into the nascent viral capsid. At late stages of infection, multiple variants of the viral genome reside within the nucleus, comprising already encapsidated genomes, genomes still involved in rolling-cycle replication and transcribed genomes. This heterogeneity makes it difficult to analyse the viral genome structure at this time point. Nonetheless, clear depletion of intrachromosomal interaction of the 7.2 kb viral intron can be observed.

While looking for host-pathogen DNA:DNA interactions, it became clear that these interactions are initiated at open chromatin in the nuclear center, followed by the expansion of VRCs and consequently leading to margination of the host chromatin to the nuclear periphery and a dramatic decrease in both nuclear and chromatin volume. Strikingly, most of these host pathogen interactions between the host and virus originated from the 7.2 kb viral intron, which was shown to have a very long half-life, late expression kinetics and is important for replication *in vivo* but has no significant function *in vitro* (Schwarz et al., 2013). In response to several different stresses, including heat and salt stress (Vilborg et al., 2015), and HSV-1 (Rutkowski et al., 2015) transcription downstream of genes for tens of kb has been reported. Vilborg et al. proposed that these downstream transcripts may help to maintain nuclear integrity after stress as they found the *doSERBP1* transcripts to remain at the site of synthesis. Thus, they might reflect *cis*-acting nuclear lncRNAs. Furthermore, it was shown that a large group of heterogeneous transcripts consisting of repetitive element-containing RNAs tethered to the nuclear scaffold surrounding chromosomes help maintain euchromatin (Hall et al., 2014). Possibly, those stress-inducible transcripts reinforce the chromosome-associated nuclear scaffold in situations where mechanical pressure forces chromatin to condense. These transcripts are part of a nuclear stress response to many different stressors (Henning & Michalski, submitted). However, we found that most read-through transcripts are released from

the chromatin into the nucleoplasm, but reside within the nucleus. I speculate that they do not have to remain in close proximity to their locus of origination to function as nuclear scaffold. A similar mechanism would be possible for the produced viral lncRNA. Furthermore, a potential mechanism of function could be similar to cellular repressive lncRNAs, such as *Xist*, which plays an instrumental role in inactivation of one of the female X chromosomes. *Xist* acts in *cis* and is directly placed on specific loci along the chromosome while being transcribed (Nora et al., 2012; Simon et al., 2013). This is only speculative and RNA-FISH experiments would provide useful insights into the nuclear localization of the RNA. Furthermore, whether the contact is mediated via the DNA or the RNA is not known. RNA affinity purification (RAP) experiments (Engreitz et al., 2015) could provide information about if and which proteins are involved, the involvement of other RNA species and could provide a detailed picture of target host genomic DNA loci.

4.5 Conclusion

In conclusion, I successfully established the Hi-C protocol to lytic mCMV infected NIH-3T3 murine fibroblasts and further enriched these Hi-C libraries for specific loci of interest. I provided strong evidence that the overall folding pattern of the host genome is highly preserved throughout the lytic infection indicating that the dramatic macroscopic changes in nuclear architecture do not result in a complete distortion of the nuclear architecture at gene level. Furthermore, the observed compaction only occurs at L1 or L2 isochores overlapping LADs, whereas the rest of the genome remains relatively more open. Additionally, I found that most promoters and their potential regulatory elements are in close 3D proximity already in non-infected cells and these contacts are maintained throughout the infection, although the activity of the potential regulatory elements is changing.

I further provided the first large DNA viral heatmaps, which reflect known features of viral genome architecture. Strikingly, I found the locus encoding for the 7.2 kb viral intron to be involved in contacts with the host, especially at the time of infection when the RNA is produced.

5 Integrated HPV 16 genomes interact with the host genome and modulate host gene expression

5.1 Introduction

Viral infections are the causing agent for approximately 15 % of all cancer cases worldwide (zur Hausen, 1991). Cervical cancer is one of the most common cancer-related mortality in women worldwide, with an estimated 266,000 woman expected to die from the disease each year (Ferlay et al., 2015). In almost all cases (99.7 %) the disease is driven by the persistent infection and ineffective clearance of HPV. As such, HPV infection is accepted to be the major cause of cervical cancer (Durst et al., 1983; Walboomers et al., 1999), especially the high risk HPVs (HRHPVs) including HPV16, 18 and 45 are associated with over 90 % of cervical malignancies, with HPV16 alone accounting for over half of all cases worldwide (Scheurer et al., 2005; Zheng et al., 2006).

Viruses have established a variety of mechanisms to function as a carcinogen, resulting in immortalisation and transformation of the infected cells. HPV 16 encodes for two oncogenic proteins, E6 and E7, which function synergistically to confer limitless replicative potential, evasion of apoptosis and genome instability (for details see 1.2.3), all of which are hallmarks of cancer (Hanahan et al., 2000). Furthermore, the cooperative action of E6 and E7 leads to the emergence of clonal cell populations with a growth advantage, predisposition for transformation and malignant progression (Hickman et al., 2002; Moody et al., 2010).

Throughout the normal life cycle the viral genome is maintained as an extrachromosomal episome, although integration of the viral genome has been shown to correlate with the progression of precancerous lesions into invasive cancers (Hu et al., 2015; Pett et al., 2007). Additionally, HPV genome integration is associated with progression from polyclonal to monoclonal status, indicating that certain integration events confer a selective advantage in a mixed cell population (Ueda et al., 2003). Remarkably, in the case of multiple concatarmeric HPV copies integrated, often the most downstream genome remains transcriptionally active, whereas the others are silenced by DNA methylation, which is thought to reflect clonal selection due to an optimal level of viral oncogene expression (Van Tine et al., 2004). Moreover, HPV integration sites are usually found at only a single or few chromosomal loci in clonal cell populations of cervical cancer (Ziegert et al., 2003). CFS of the human genome have been shown to be hotspots of HPV integration (Stanley et al., 2006; Thorland et al., 2003; Thorland et al., 2000). The expression of the proto-oncogene *c-MYC*, encoded in one of those regions, is maybe increased as a result of HPV integration near that site (Ojesina et al., 2014; Schmitz et al., 2012). ChIA-PET data demonstrated a long-range *cis* interaction, spanning more than 500 kb, between the integrated HPV18 promoter/enhancer and the *MYC* gene

(Adey et al., 2013). The expression of host genes at or near the integration site is changed as a result of HPV integration, which has been shown to cause a wide-range of somatic mutations, copy number variations and structural rearrangements of the host genome (Burk, 2017). As well as increased expression of the viral oncogenes, it is likely that alterations to host gene expression may also promote malignant progression and are possibly regulated by genomic re-arrangements due to viral integrations.

5.2 Objectives and outline

The W12 system provide a unique opportunity to study virus host DNA-DNA interactions in humans cells and how this might lead to an advantage in growth and potentially transformation. The association between integration of the HPV16 genome into the host and the severity of disease has been widely commented on, and is known to be the major risk factor associated with disease progression.

To address this aim I have used a panel of W12 integrant clones (Pett et al., 2004) that exhibit different levels of HPV16 oncogene expression per template and have less than four copies of the virus genome to compare HPV mediated changes in host gene expression and explore the role of virus host genomic interactions thereof.

In the first part of this chapter, I confirm the quality of the generated Hi-C libraries and will further outline how SCRiBL has been successfully adapted to enrich for the HPV16 genome in the W12 integrant clone system. The need for significant enrichment of Hi-C libraries is particularly relevant due to the relatively small size of the HPV genome in comparison to the host. The successful generation of SCRiBL libraries will enable a much broader analysis of the epigenetic regulation of viral transcription in the W12 clones.

In the second part, data obtained from Hi-C and SCRiBL libraries will elucidate whether characteristic virus and host changes, such as long-range 3D interactions and resultant host gene expression changes, are a typical feature of all HPV integrants or whether they are restricted to cells with a selective growth advantage. Furthermore, we have been able to pinpoint the precise locations of HPV16 integration sites within our W12 integrant clones by sequence capture of the viral genome, coinciding with areas of open chromatin, as well as determining that integration likely occurs through microhomology mediated repair mechanisms.

5.3 Results

5.3.1 Successful generation of Hi-C libraries and sequence capture enrichment

Work presented in this chapter has been done I collaboration with Nick Coleman`s group at the Department of Pathology in Cambridge, UK, in particular with Emma Knight and Ian Groves. In order to address the above questions, we generated “in nucleus ligation” Hi-C

libraries (Nagano et al., 2013; Rao et al., 2014; Sofueva et al., 2013) in biological replicates from five different W12 integrant clones with viral genome copy numbers less than four: F, A5, D2, H and G2 (Table 5.1). We used a 4-cutter restriction enzyme, MboI, for Hi-C library generation, to maximise the resolution, especially on the viral genome. Furthermore, we generated biotinylated RNA baits against HPV16 genomic MboI restriction fragment ends to enrich those Hi-C libraries for viral genome containing di-tags. Moreover, biotinylated baits against the entire viral genome were generated to determine the viral integration site with nucleotide resolution.

As for the lytically infected mouse libraries (Chapter 4), ligation efficiency tests and the detection of known long and short range interactions by PCR, are an excellent indication of the libraries quality prior to sequencing. As such, forward primers were designed across the *RPL13A* host genomic locus to detect short-range interactions. The locus was divided based on MboI fragments, which were labelled alphabetically with fragment A contacting the TSS. PCR reactions with the primer pair B:G gave the expected product of 319 bp in the parental NCx libraries and the two HPV-infected cell libraries, namely F and G2, but not on genomic DNA in any of the cases, indicating successful digestion and subsequent ligation (Figure 5.1a). Additional PCR products of varying size were detected by agarose electrophoresis, most likely representing the concatemeric nature of Hi-C material after the ligation and prior to sonication. For robustness, an alternative PCR method for detecting short-range interactions in Hi-C libraries was utilised. Therefore, a forward primer, spanning two restriction fragments (D and J of the *RPL13A* locus) and containing the 5`-GATCGATC-3` sequence, which was generated by MboI restriction digestion, fill-in and subsequent blunt-end ligation, was designed. The corresponding reverse primer resided in fragment J, which matches the fragment of the 3`-end of the forward primer. After annealing temperature optimisation, the desired 401 bp product could only be detected in Hi-C libraries but not on genomic DNA for all Hi-C libraries generated, further suggesting successful digestion and re-ligation (Figure 5.1b). These tests do not reveal any information about the fill-in, but as for BglII digestion (Chapter 3), fragmentation with MboI and subsequent fill-in followed by blunt-end ligation creates a novel ClaI site, although in the case of MboI the initial site is not lost. The PCR products generated by using the two forward primers (B:G) was subjected to ClaI digestion for all libraries and for all of the samples the 319 bp product was almost completely cut into a 219 bp and a 100 bp band (Figure 5.1c). The ligation efficiency was estimated by quantifying the intensity of the cut and uncut bands using Image J software. The ligation efficiency was high in all W12 Hi-C libraries ranging from 71.1 % (W12 clone A5 rep I) to 94.2 % (W12 clone G2 rep I); this was also suggestive of successful library generation.

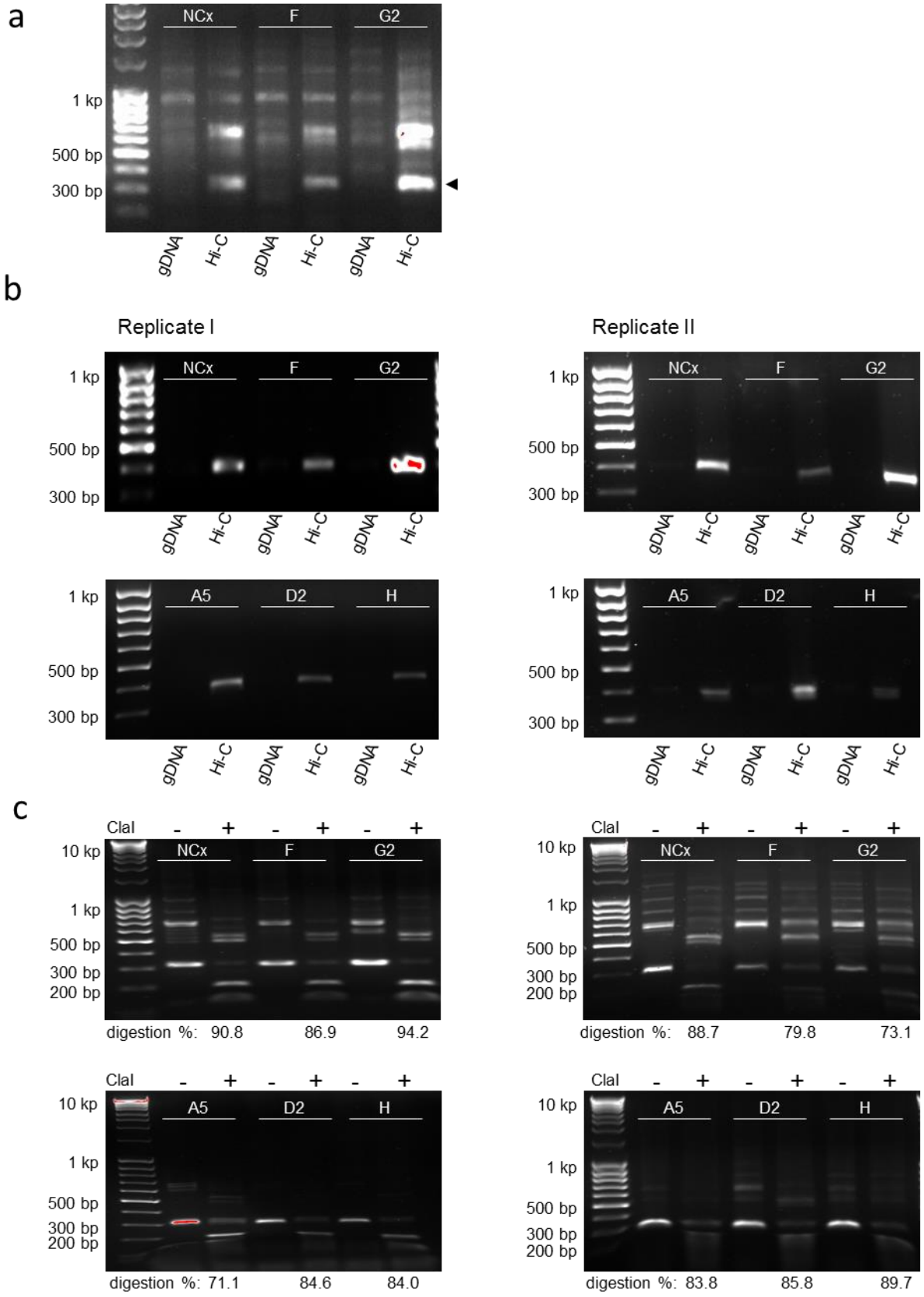


Figure 5.1 | W12 clone Hi-C library preparation quality controls

(a) PCR reaction products of NCx, W12 F and W12 G2. For each derived from genomic and Hi-C library material using two forward primers on MboI fragments B and G in the RPL13A locus, with fragment A containing the TSS

and then numbered alphabetically. Black arrowheads indicate the B:G specific PCR product, which is only present in Hi-C material**. (b)** Additionally, a forward primer, spanning a non-consecutive restriction site was used in combination with a reverse primer on the landing MboI fragment to specifically amplify a known short-range interaction. The Hi-C specific product is present in all Hi-C libraries in both replicates, but not in the genomic DNA libraries, indicating successful digestion and ligation of the Hi-C libraries. **(c)** Restriction digestion with MboI, followed by fill-in and blunt end ligation creates a new ClaI restriction site, which can be used to determine the Hi-C ligation efficiency. Therefore, the short-range product amplified in the RPL13A locus using primer pairs B and G from Hi-C libraries, was either not digested or digested with ClaI for all 12 Hi-C libraries generated. For each clone, a high digestion efficiency was observed.

The optimal number of PCR cycles for the final amplification of the Hi-C material was determined by running aliquots (1/20th library) with different amplification cycles and visualisation of the products following agarose gel electrophoresis (Figure 5.2). The aim of this test is to avoid over-amplification of the libraries, but to generate enough material for sequencing and region capture. The necessary number of PCR amplification cycles for Hi-C libraries was decided by choosing one fewer number of cycles than that at which a smear was visible; this was constant across samples generated at the same time, i.e. NCx, F and G2 = 8 cycles and A5, D2 and H = 9 cycles.

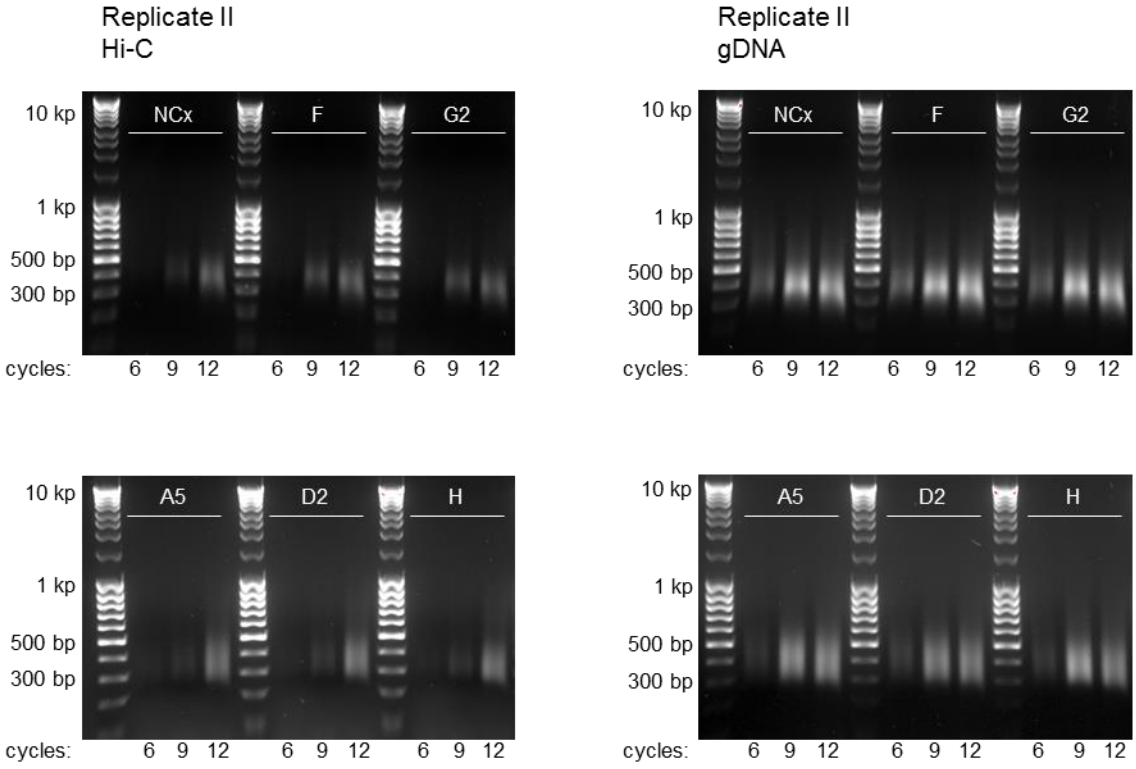


Figure 5.2 | PCR test amplification f Hi-C and genomic libraries

PCR tests were performed on 1/20th of the library material to determine the optimal number of PCR cycles for the final PCR amplification. The desired amount of final library lies around 1 µg, which is enough for capturing and NGS. The number chosen for final library amplification is shown under the gels.

The size distribution and concentration of the final Hi-C libraries were assessed by Bioanalyzer analysis (data not shown). Libraries were found to be in the optimal range of 300-700 bp for Illumina sequencing, demonstrating precise sonication and size selection. All libraries showed sufficient but not over amplification. Hi-C libraries from clones W12 G2 and W12 D2 were each sequenced with both replicates on one HiSeq 2500 lane with a 50 bp paired-end setup. All libraries were subjected to sequence capture enrichment (SCRiBL).

Table 5.1 | Details of the W12 clones studied

clone	ploidy	integration site	copy number E6	Expression E6 per template	category
F	2N	4q13.3	1	248.6	high
A5	2N	8q11.21	1	215.6	high
D2	2N	18q21.2	3	118.5	medium
H	2N	4q21.23	1	100.1	medium
G2	2N	21q22.1	3	37.5	low

A key experimental aim was to accurately identify the virus-host breakpoint junctions in the W12 clones. To this end, enrichment of the HPV16 genome sequence from the genomic libraries, which had been fragmented by sonication only, was first used. The pSP64 plasmid, in which the W12E HPV16 genome is cloned (Cinzia Scarpini) was used to generate fragments of HPV16 DNA that were later *in vitro* transcribed for use in the hybridisation reaction with the W12 clone genomic libraries. Four sets of primers, containing the T7 promoter sequence at the 5` end, were designed to evenly cover the entire HPV16 W12E genome (see Table 5.2).

Table 5.2 | DNA primers spanning W12E genome for whole HPV16 genome capture

name	HPV16 coordinates	product length (bp)
HPV_DNA_block1	4-2,004	2,000
HPV_DNA_block2	1,985-3,941	1,956
HPV_DNA_block3	3,903-5,852	1,950
HPV_DNA_block4	5,826-7,891	2,066

PCR products (Figure 5.3a) of amplified HPV16 genome were kept separate or were combined in equimolar amounts and subjected to in vitro transcription using biotin-UTP. Despite a smear for all but HPV_DNA_block3, which was not loaded properly, the full-length product can be seen in all of the reactions (Figure 5.3b). Chemical fragmentation was used to generate biotinylated RNA oligonucleotides of approximately 150 nt complementary to the HPV16 genome (Figure 5.3c). Of note, because of the observed smear, prior to the *in vitro* transcription, this approach is not quantitative but only qualitative, which is sufficient for determining the host break points.

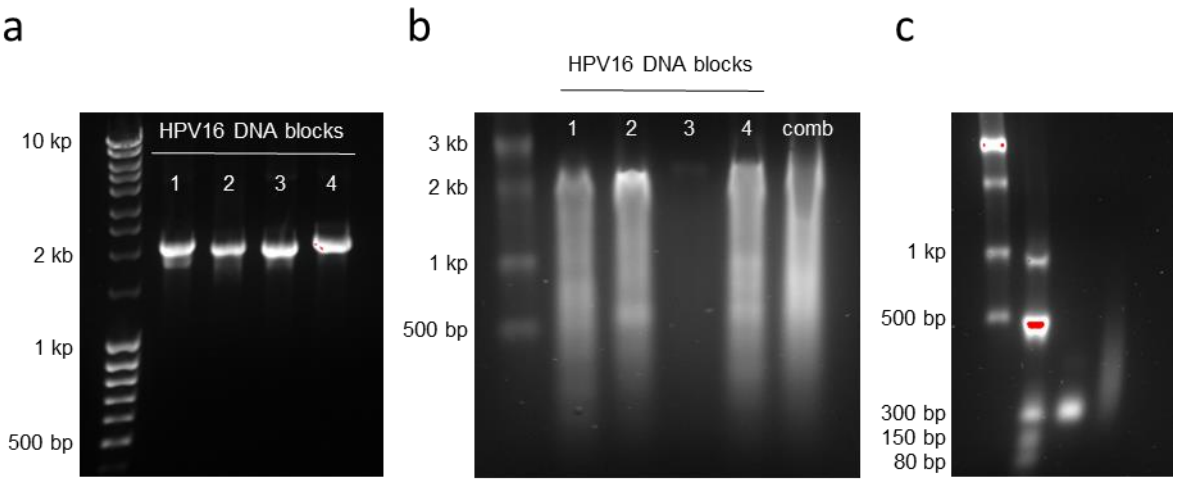


Figure 5.3 | Genomic sequence capture bait generation

(a) The HPV16 genome was PCR amplified from the BAC pSP64 HPV16, using four consecutive slightly overlapping PCR primer pairs, where each of the four forward primers contained the T7 promoter sequence, enabling in vitro transcription of the four templates using biotin-UTP. Expected DNA template lengths are 2000 bp, 1956 bp, 1950 bp and 2066 bp. **(b)** RNA product derived from *in vitro* transcription of the four DNA blocks covering the entire HPV16 genome. Individual products and equimolar pool are depicted as indicated. **(c)** Obtained RNA was chemically fragmented to a size of around 130 nt. The first non-ladder lane depicts reference biotinylated RNA at a size of 130 nt, whereas the second lane contains the fragmented biotinylated HPV16 RNA baits.

To identify 3D interactions between genomes of the integrated virus and the host, it was necessary to enrich the W12 clone Hi-C libraries for the HPV16 genome, with biotinylated RNA baits capturing the restriction fragment ends of the MboI fragments originating from the viral sequence. We used a gBlock® Gene Fragments from Integrated DNA Technologies (IDT) based approach, which is illustrated in Figure 5.4. First, ~120 bp long sequences of DNA complementary to the 5'-end of HPV16 genomic MboI fragments were constructed into gBlocks® flanked by two restriction enzymes, either BglII/HindIII (gBlock1) or BglIII/SpeI (gBlock2). The gBlocks were cloned using the Zero-Blunt® TOPO® cloning and the 1.2 kb and 1 kb products of gBlock1 and gBlock2, respectively, were isolated and gel purified (Figure 5.5a). Double digestion with the appropriate restriction enzymes released the ~120 bp DNA sequences and allowed for site specific ligation of T7 promoter adapters (Figure 5.5b; 180 bp band) followed by in vitro transcription using biotinylated-UTP. A tight band of 130 nt was observed (Figure 5.5c).

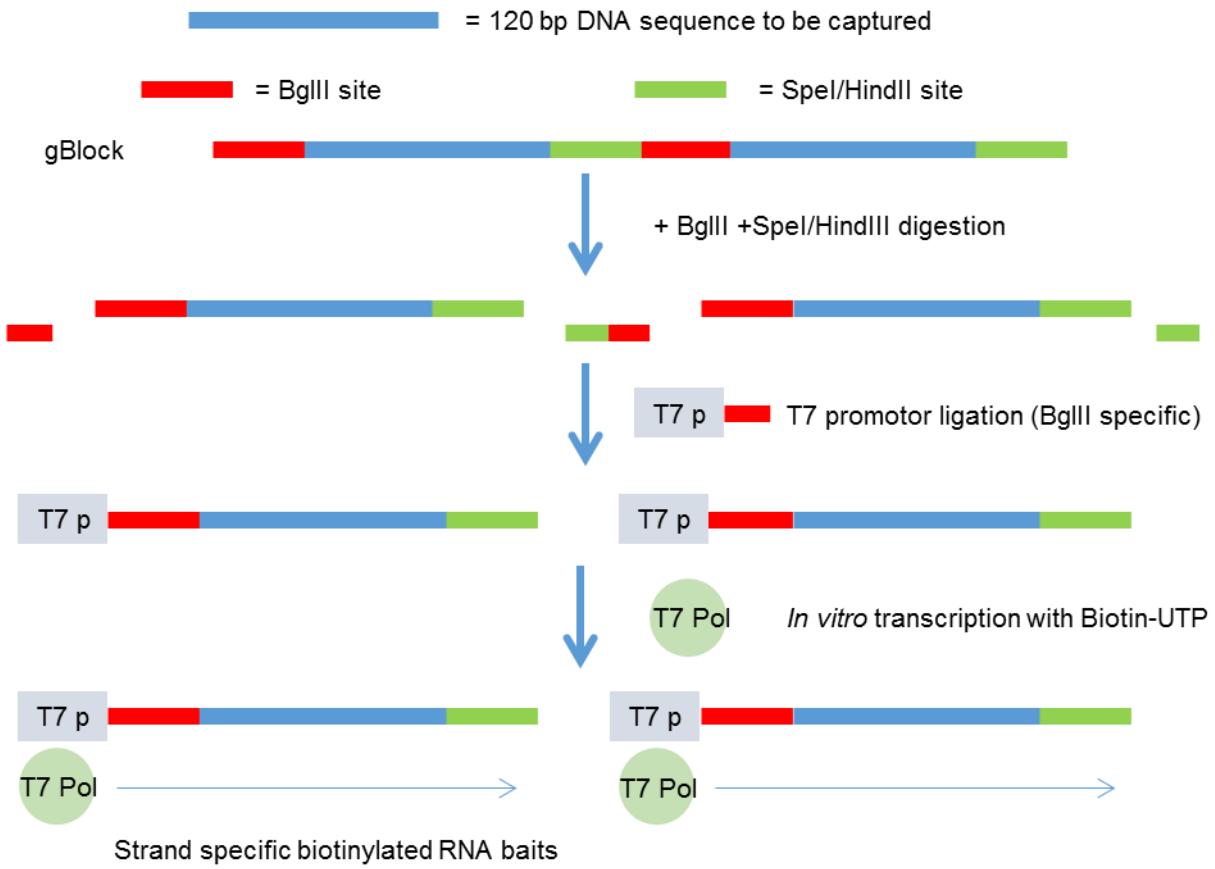


Figure 5.4 | Schematic of the gBlock based approach for HPV16 capture from Hi-C

120 bp DNA sequences, covering the HPV16 MboI restriction fragment ends, were encompassed by specific restriction fragment sites, in such a way that on the 5` end there was a BglII site, whereas on the 3` end there was either a HindIII (block1) of a SpeI site (block2). This allowed for BglII specific T7 promoter adapter (carrying the BamHI overhang) ligation, in the presence of BglII and SpeI/HindIII. Subsequent in vitro transcription using biotin-UTP resulted in 130 nt biotinylated RNA baits. Sequences of the two gBlocks can be found in the appendix.

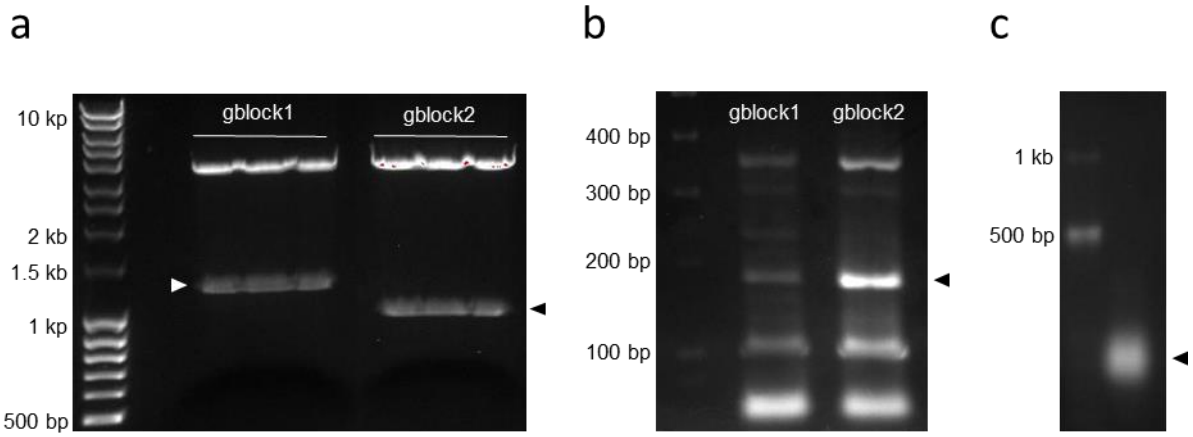


Figure 5.5 | Quality control of SCRiBL bait generation

(a) gBlocks were cloned into Zero-Blunt® TOPO® cloning vector and isolated by EcoRI digestion, followed by gelelectrophoresis and gel extraction. **(b)** gBlocks were digested with BglII and either HindIII or SpeI to release the fragments with the desired sticky ends. This was followed by the ligation of pre-annealed T7 promoter adapters in

the presence of BglII and HindIII/SpeI. The arrowhead is indicating the desired ligation products of T7 promoter adapter ligated to the 130 bp fragments. **(c)** Equimolar amounts of each gBlock fragments were combined, *in vitro* transcribed using biotin-UTP, purified and the integrity was checked by agarose gelelectrophoresis. The arrowhead indicates the desired product at 130 nt.

After the hybridisation reaction and streptavidin pull down on RNA/DNA hybrid complexes a test PCR was carried out to determine the number of PCR cycles required to generate enough material for genomic sequencing without the introduction of too many PCR duplicates. A smear of 300-800 bp was produced after 12 amplification cycles, although these were very faint in the NCx, F and G2 replicates (Figure 5.6a). The final number of amplification cycles chosen was one directly below that at which a smear was visible; hence the final F and G2 libraries had an additional amplification cycle compared with A5, D2 and H (11 vs. 10 cycles, respectively). The PCR conditions for the captured undigested libraries were equally determined using PE primers 1.0 and 2.0 (Illumina) in test PCR on 1/20th aliquots (Figure 5.6b). Both, the Hi-C and undigested NCx libraries, were not sequenced and a final library not generated; hence the amplification cycle number determination was not necessary.

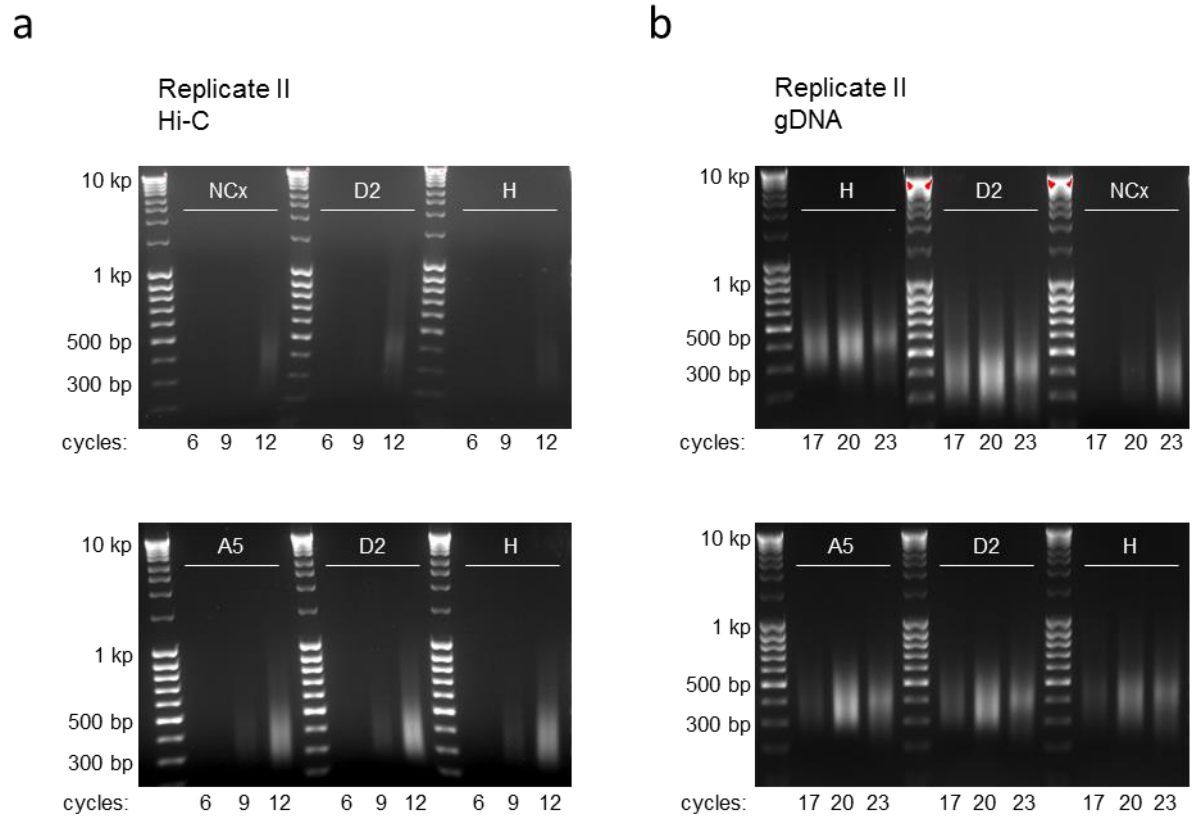


Figure 5.6 | SCRiBL and capture-Seq test PCRs

A 1/20th of the post-capture libraries was amplified using varying numbers of PCR cycles to determine the optimal number of PCR cycles for the final PCR amplification. SCRiBL libraries derived from clones F and G2 were amplified with 11 PCR cycles, whereas SCRiBL libraries from clones A, D2 and H were amplified using 10 PCR cycles. All,

but D2, captured genomic libraries were amplified with 15 PCR cycles. Captured genomic DNA from clone D2 was amplified with 16 cycles.

Most bioinformatic analysis including the mapping and aligning of sequence reads was carried out in collaboration with Jack Monahan (EBI-EMBL). After completion of the 50 bp, paired-end sequencing run of each of the SCRiBL and Hi-C libraries, mapping and quality control checks were conducted using HiCUP (Wingett et al., 2015). As before, HiCUP maps the data to a reference genome, in this case human genome GRCh38, including the HPV16 genome as an extra chromosome, and filters out experimental artefacts and PCR duplicates.

All but one, namely D2 replicate I, of the Hi-C libraries showed a high number of valid read pairs. Furthermore, a good indicator of the libraries` quality is the *cis/trans* ratio within the valid pairs. This is only informative for Hi-C libraries, as enriched libraries have a strong bias towards the captured regions, in our case the viral genome, which was included in the reference genome in *trans*. All sequenced Hi-C libraries showed less than 10 % *trans* reads Figure 5.7b), which in agreement with previously published libraries and indicates high quality (Nagano et al., 2015). For the SCRiBL libraries, the percentage of total mapped reads was fairly consistent across all of the libraries (25.7-39.9 %), however the percentage of valid reads was much higher in SCRiBL libraries of D2 rep II, G2 repI and G2 repII (24.9, 34.0 and 26.1 %, respectively) compared to the 16.0 % average of the seven other libraries (Figure 5.7c). Of the invalid read pairs, re-ligation events (8.1-27.5 %) and dangling ends (2.7-52.0 %) were the most common cause of invalidity. After HiCUP processing, an additional Perl script was run on the SCRiBL libraries to remove off-target read-pairs, were none of the two reads from a pair map to the viral genome.

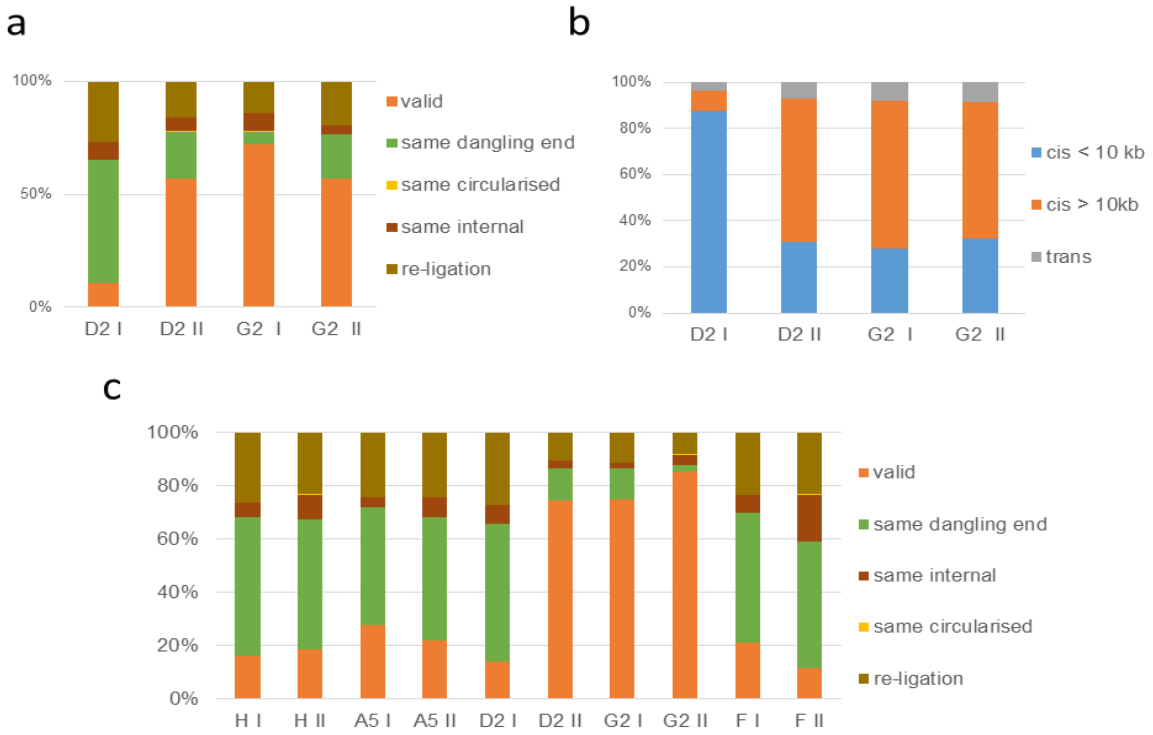


Figure 5.7 | HiCUP statistics of sequenced SCRiBL and Hi-C libraries

Only Hi-C di-tags were both ends are originating from restriction fragment ends coming from different fragments contain useful information about the 3D organisation, the rest are filtered out by the HiCUP pipeline (Wingett et al., 2015). 100 % stacked column chart showing the percentage of valid and invalid read pairs, compared to the total number of mapped reads for **(a)** the Hi-C libraries and **(c)** the SCRiBL libraries. **(b)** 100 % stacked bar chart displaying the percentage of close *cis* (<10 kb), far *cis* (>10 kb) and *trans* reads for the valid read-pairs in each of the generated Hi-C libraries.

Table 5.3 | Obtained Hi-C sequencing read numbers

	D2 I	D2 II	G2 I	G2 II
valid	2,822,499	16,805,551	31,469,583	21,987,034
invalid	23,875,491	12,724,446	12,210,715	16,606,201
cis < 10 kb	2,405,568	5,123,188	8,354,533	6,988,097
cis > 10 kb	248,188	10,196,575	19,025,143	12,771,030
trans	98,398	1,204,721	2,391,703	1,855,665

Table 5.4 | Obtained SCRiBL sequencing read numbers

	H I	H II	A5 I	A5 II	D2 I	D2 II
paired	21,909,789	18,468,258	18,877,049	17,884,689	23,310,201	23,111,170
valid	3,496,521	3,416,705	5,262,234	3,904,936	3,247,691	17,204,769
invalid	18,413,268	15,051,553	13,614,815	13,979,753	20,062,510	5,906,401
same circularised	4,503	7,183	13,883	7,627	2,651	42,023
same dangling end	11,429,535	8,995,563	8,309,537	8,277,939	12,064,098	2,783,555
same internal	1,231,083	1,753,222	731,997	1,370,084	1,617,965	660,006
re-ligation	5,748,147	4,295,585	4,559,398	4,324,103	6,377,796	2,420,817

	G2 I	G2 II	F I	F II
paired	21,507,596	39,169,354	22,350,624	28,885,328
valid	16,108,584	33,384,729	4,728,217	3,337,875
invalid	5,399,012	5,784,625	17,622,407	25,547,453
same circularised	41,093	103,457	9,492	5,490
same dangling end	2,498,283	1,038,276	10,838,740	13,700,429
same internal	410,512	1,480,308	1,492,107	5,119,714
re-ligation	2,449,124	3,162,584	5,282,068	6,721,820

Taken together, these results show that the “in nucleus ligation” Hi-C protocol has successfully been adjusted to the use of a 4-cutter restriction enzyme in human keratinocyte derived cell lines. Furthermore, we established a protocol to enrich for the viral genome in Hi-C data and from genomic DNA.

5.3.2 Integrated HPV16 genomes interact with the host genome

To investigate which regions of the integrated HPV16 genomes are in close spatial proximity with the host genome, loci on the human genome significantly interacting more frequently with the viral genome than the rest of the genome, were identified using GOTHiC (Mifsud et al., 2017) and visualised using Circos plots. In each panel (Figure 5.8a-e), a single line represents a significant virus host interaction and is coloured according to the viral gene it is originating from. The frequency of *cis* interactions between the virus and the host is known to be greatest for host sequences at the site of integration and to decrease with distance (Lajoie et al., 2015). Hence, analysis of the SCRiBL Hi-C data simultaneously enabled detection of the HPV16 integration locus in each of the clones. For all clones a single integration site, determined by a single interacting locus on a specific chromosome, was detected. For all clones, interactions originated from the vicinity of restriction fragments ends on the viral genome, which suggests the successful generation and enrichment of Hi-C di-tags. For the W12 clone G2 the greatest percentage originated from the viral fragment containing the E7 oncogene. The integrated HPV16 genome is exclusively interacting with chromosome 5 (Figure 5.8a). The number of virus-host reads that were captured and mapped for W12 clones D2 (84.3 %), F (11.2 %), A5 (12.4 %) and H (10.3 %) were lower than for clone G2 arbitrarily set at 100 %. As a result, there are reduced numbers of virus-host reads in the circos plot analysis. The HPV16 genome in clone D2 has most likely also integrated into chromosome 5, but in a completely different locus, with most of the host-pathogen interactions formed by the 5` half of the viral L1 gene (Figure 5.8b). For clone H, the captured reads indicate that interactions are mainly mediated by the early genes E6, E7, E2 and L1, with the majority coming from E2 and are uniquely formed with one locus on chromosome 4 (Figure 5.8c). Strikingly, we found that W12 clones F and A5 had

the same integration site, with virus-host reads converging to the same region of chromosome 4 (Figure 5.8d and e), with the majority of interactions being mediated via the viral gene E2 in both cases. In all the W12 integrant clones tested, the HPV16 genome was shown to interact with regions of host chromosomes in *cis*; there were no examples of the virus interacting in *trans*.

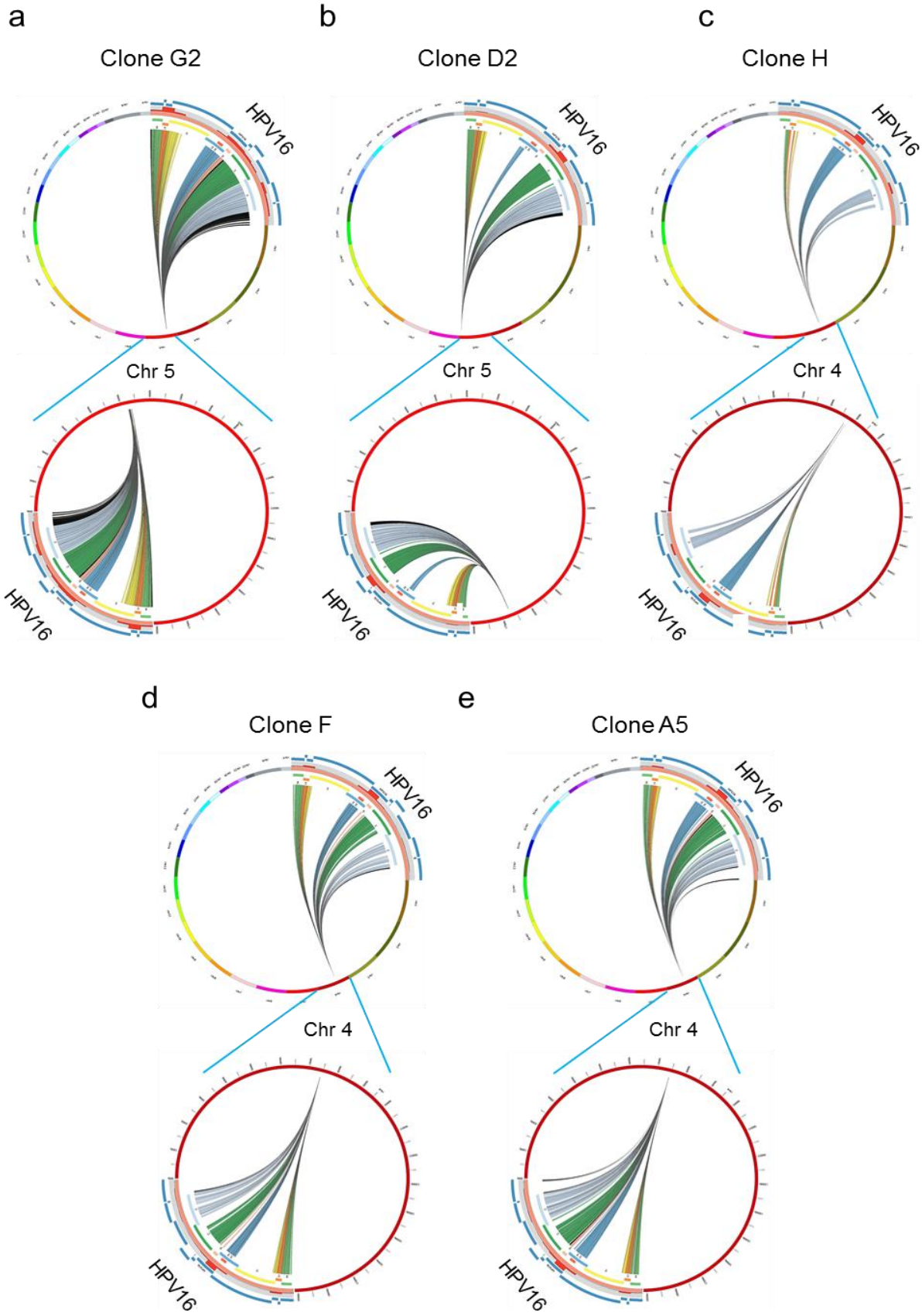


Figure 5.8 | Circos plots indicating 3D interaction between the integrated HPV16 genomes and the host genome

Each of the lines within a circle represents a significant virus-host read, determined by GOTHiC. Circos plots are depicted for the higher quality replicate for each of the clones as indicated: **(a)** W12 G2 replicate II, **(b)** W12 D2 replicate I, **(c)** W12 H replicate I, **(d)** W12 F replicate II and **(e)** W12 A5 replicate I. Reads were coloured based on the viral gene they were originating from: E6 = green, E7 = orange, E1 = yellow, E2 = blue, E4 = red, E5 = pink, L1 = dark green, L2 = light blue and the LCR = black. The top circus plot of each panel comprise the HPV1 6 genome (orange) and the entire host genome, with individual chromosomes labelled with different colours. The bottom circus pot of each of the panels comprise the HPV16 genome (orange) and the single host chromosome where 3D interactions occur. The histogram in red, which is split into 500 bp windows, is indicating the percentage of reads originating from that locus on the viral genome.

5.3.3 Virus-host breakpoint identification at nucleotide resolution.

Capturing the viral genome from sonicated genomic material enabled us to determine the virus-host breakpoints at nucleotide resolution by sequencing the chimeric sequences. Analysis of the capture-Seq experiment determined that for each clone, peaks of virus-host reads mapped to two distinct sites of the host genome, independent of the viral copy number. These findings were validated by qPCR and Sanger sequencing (Emma Knight, data not shown). In the clone G2, the viral genome is linearised by breaking the viral genome in the *E2* ORF [5`: 2940 and 3`: 2,768] resulting in a loss of 173 bp and placing the majority of the *E2* gene upstream of the viral early promoter. Additionally, the three copies of the HPV16 genome integrate into chromosome 5 in an intergenic region of the host genome [5`: 52,681,626 and 3`: 52,655,805]. The four HPV16 genomes in clone D2 are also integrated into chromosome 5 [5': 167,112,984 and 3': 167,141,612]. Linearisation of the HPV16 genome occurs via breakage in the *L2* [5': 4,361] and *E2* [3': 3,272] ORFs, resulting in a 1,089 bp deletion of the virus genome. Furthermore, the orientation of the virus promoter opposes that of the transcription of the host gene *TENM2* into which the HPV16 genome has integrated. In W12 clone H only a single copy of the HPV16 genome has integrated into chromosome 4 within the host gene *MAPK10*; moreover transcription from the virus early promoter occurs in the same direction as transcription of the host gene. Virus integration results in a large deletion of the host genome, with the 5' and 3' host breakpoints separated by more than 170 kb [5':86,983,196 and 3':87,153,458]. In addition, in comparison with the other W12 integrant clones included in this study, a large proportion of the virus genome is also deleted; the HPV16 genome is broken in the *L1* [5':5,883] and *E2* [3': 3,751] ORFs, meaning that the length of the integrated virus is 5,773 bp. The Sanger sequencing data confirmed our initial finding that clone F and A5 have the same site of HPV16 integration. In both clones, a single copy of the virus integrates into chromosome 4 within the host gene *RASSF6* [5':74,549,681 and 3':74,480,662]. Sanger sequencing also revealed an 18 bp truncation and rearrangement of the HPV16 genome at the 3' breakpoint whereby a region of 54 bp (3,637 to 3,690) is inverted. The linearisation of the HPV16 genome and resultant rearrangements occur within the *E2* ORF [5':3,677 and 3':3,637] and as a result places a portion of the *E2* ORF upstream of the virus early promoter.

With the exact virus-host breakpoints in hand, we were able to identify microhomologies around the integration sites (discussed elsewhere; Knight et al. in preparation). Furthermore, we were able to identify the chromatin statues of the 5 Mb around the integration sites based on publicly available datasets derived from normal human epidermal keratinocytes (NHEKL, ENCODE). In G2, the viral genome has integrated into an open, active enhancer region, determined by Dnase I, H3K4me1 and H3K27Ac peaks. Active marks can also be found at the integration site in the clones F, A5 and D2. In H, the viral genome has integrated into an accessible region without any particular activating histone marks. Strikingly, all integration sites are depleted for repressive marks, such as H3K27me3.

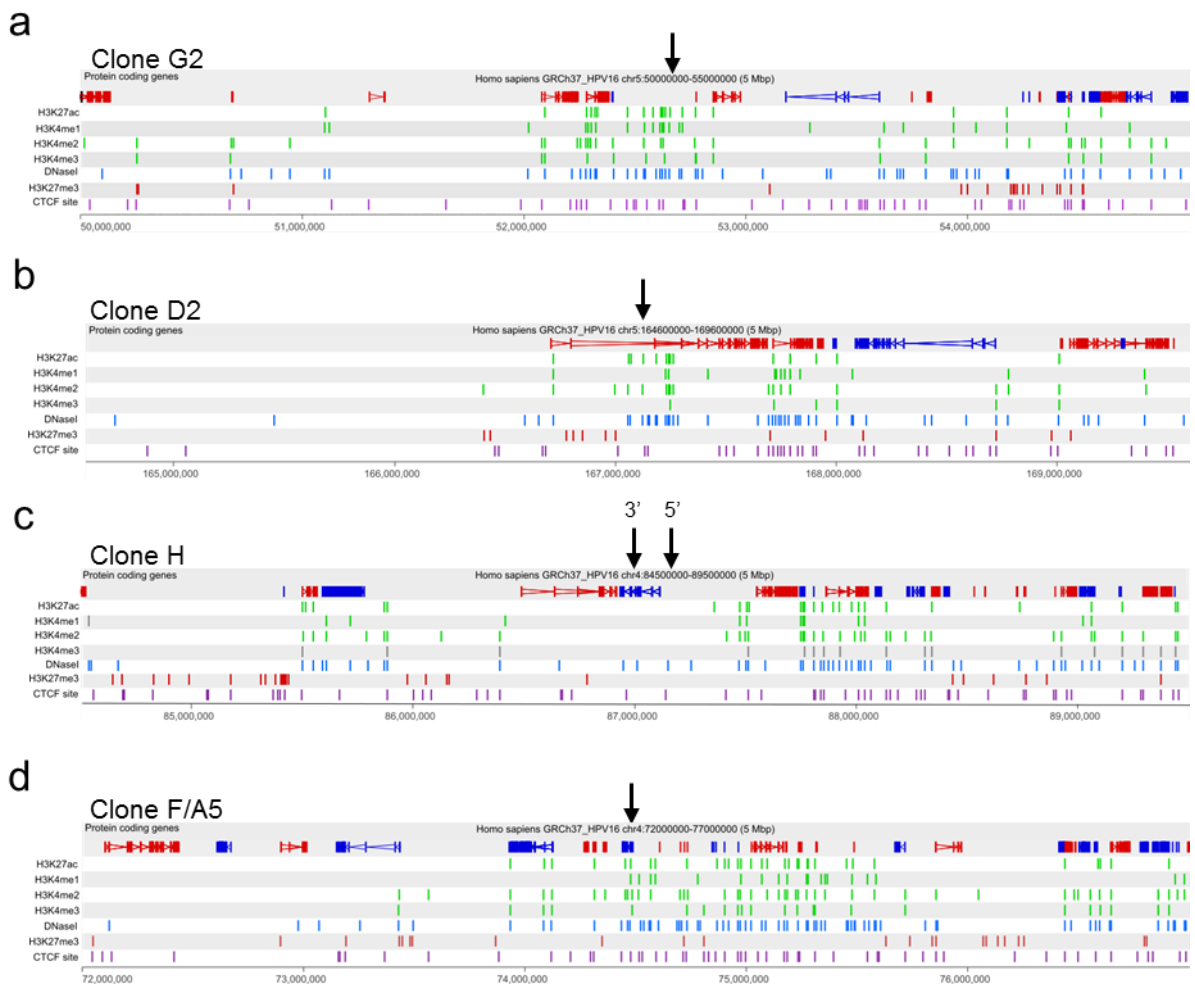


Figure 5.9 | Chromatin marks around HPV16 integration sites

Each panel shows the 5 Mb host genomic region surrounding the viral integration site (black arrow; 5` and 3` separated in clone H due to deletion). Protein coding regions are shown and colour coded based on orientation (red = + strand; blue = - strand). ChIP data derived from normal human epidermal keratinocytes (NHEK) is aligned with the host genome (taken from ENCODE). Activating histone marks, such as H3K4me1/2/3 and H3K27Ac are coloured in green; DNAseI hypersensitivity sites are coloured in blue; the repressive H3K27me3 marks is coloured in red and CTCF binding sites are coloured in purple. **(a)** W12 G2, **(b)** W12 D2 I, **(c)** W12 H **(d)** W12 F/A5.

These data show that capturing the viral genome in Hi-C data and from sonicated genomic DNA can be utilised to identify the integration sites with very high resolution. Furthermore, no *trans* interactions between the viral genome and other chromosomes than the one the virus has integrated in could be observed, restraining the viral epigenomic regulation to a more local environment.

5.3.4 Short- and long-range 3D interactions occur between the HPV16 and host genomes regardless of cell selection during early cervical carcinogenesis

To assess whether there are specific short- and/or long-range contacts originating from the integrated viral genomes, the SCRiBL data was visualised using Seqmonk. Visual inspection of v4C profiles obtained from the entire viral genomes revealed multiple short and long range interactions across the 5 clones. Close analysis of the HPV16 integration locus in all clones at high normalised read depth revealed clearly defined peaks that matched the integration breakpoints identified by Sanger sequencing (Figure 5.10; Figure 5.11a and Figure 5.13a).

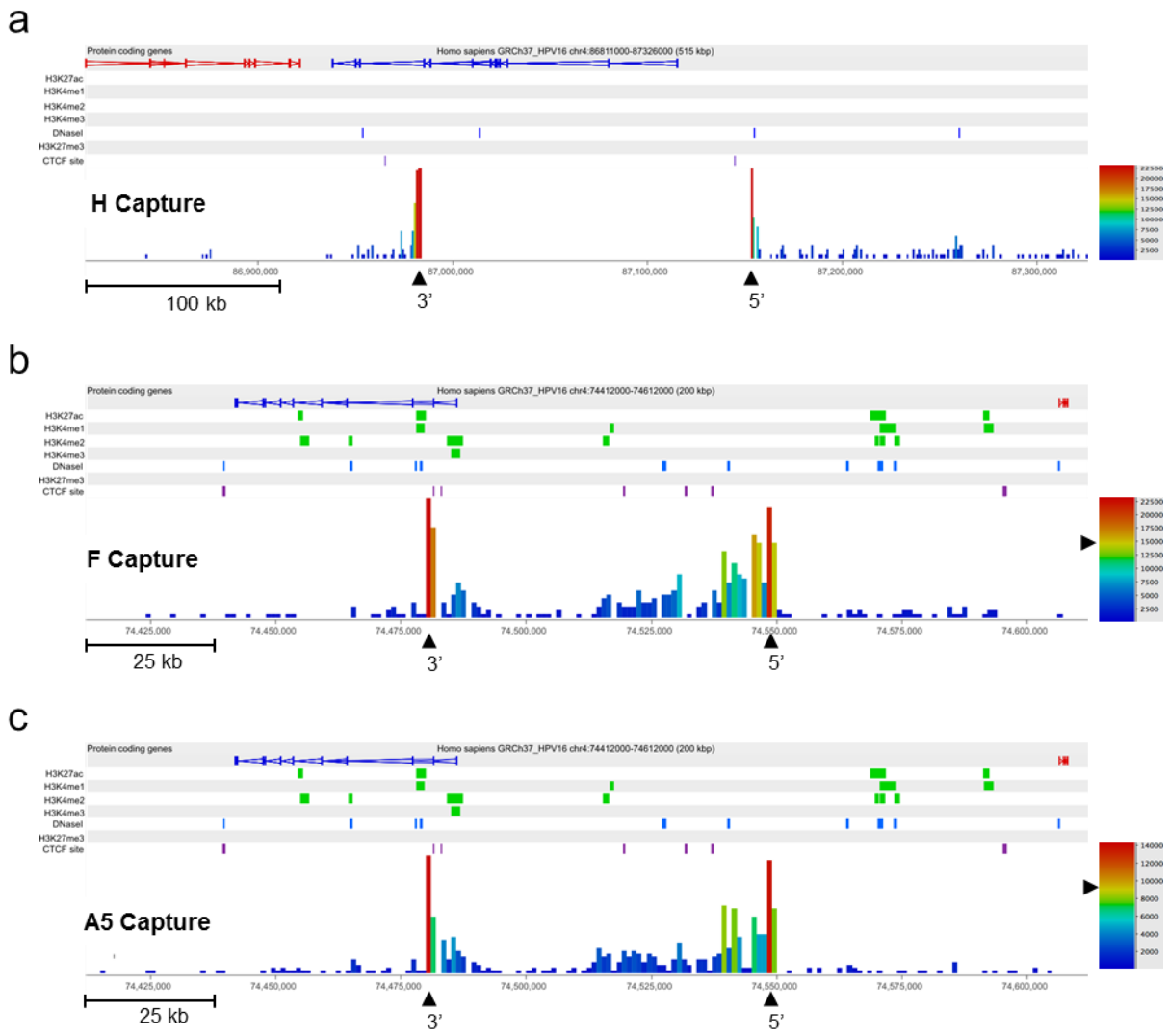


Figure 5.10 | Virus-hots breakpoints identified by SCRiBL in W12 clones H, F and A5

SCRiBL data at 1 kb resolution, across regions surrounding the viral integration sites of **(a)** W12 H, **(b)** W12 F and **(c)** W12 A5. The highest red bars are indicative of the 5` and the 3` host-virus breakpoints and are labelled with arrowheads, while the scale bar indicates the normalised read count. Additionally, protein coding regions are shown and coloured based on their orientation (red = + strand; blue = - strand). Histone marks correlating with gene activation, such as H3K4me1/2/3 and H3K27Ac are coloured in green; DNAseI hypersensitivity sites are coloured in blue; the repressive H3K27me3 marks is coloured in red and CTCF binding sites are coloured in purple.

No other interaction originating from the viral genome towards the host genome were observed for the clones A5, F and H. Nonetheless, by increasing the window size around the integration locus in clone G2 (to 700 kb) and decreasing the normalised read depth, additional multiple short- to medium-range (<500 kb) 3D interactions between the integrated virus and the surrounding host genome were detected; with distances from 34-238 kb (Figure 5.11b). Notably, the interaction peaks in clone G2 appeared to align with CTCF sites, with the majority additionally overlapping marks of enhancer regions (H3K27ac and H3K4me1) and regions of DNaseI hypersensitivity (Figure 5.11b). Expanding the window further (5 Mb) illustrated a number of long-range (>500 kb) 3D interactions between the virus and the host. The furthest and most prominent peak was located at around 53,520,000, within the first intron of host gene *ARL15*, approximately 900 kb from the HPV16 integration site (Figure 5.11c).

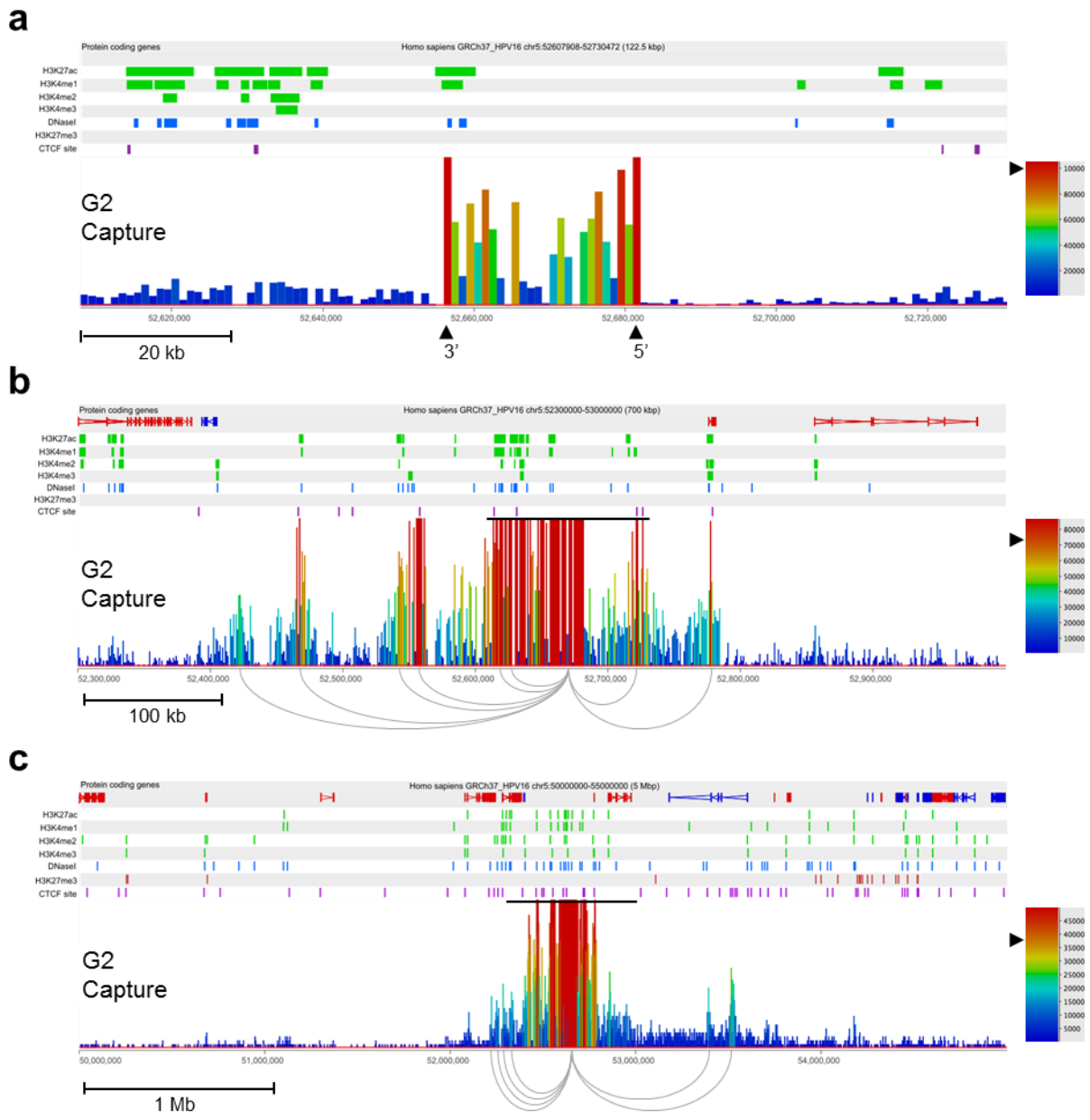


Figure 5.11 | Identification of short and long-range interactions between integrated HPV16 and the host genome in W12 clone G2

(a) SCRiBL data across a 122.5 kb region surrounding the G2 integration site. The 5` and the 3` integration sites are reflected by the tallest bars and were labelled with arrowheads. **(b)** Region of 700 kb across the viral site in clone G2 showing SCRiBL data. The black line seen above the locus indicates the genomic region seen in panel a. Arches are drawn for interactions, with more than 44,000 normalised reads. **(c)** SCRiBL data of 5 Mb across the viral integration site in clone G2. The black line seen above the locus indicates the genomic region seen in the panel above. Arches are drawn for interactions, with more than 16,000 normalised reads. In each of the panels, the key indicates the normalised read counts. Additionally, protein coding regions are shown and coloured based on their orientation (red = + strand; blue = - strand) and activating histone modifications are shown in green whereas repressive H3K27me3 is depicted I red; CTCF binding site are indicated in purple and DNaseI sites are shown in blue.

To verify and validate the long-range interaction in G2 between the integrated viral genome and the *ARL15* intron 900 kb away, detected by SCRiBL, 3D FISH was performed (by Emma Knight). Three fluorescent DNA probes were generated to hybridise to either the HPV16 genome, the first intron of *ARL15*, or a control region, matching the linear distance, but in the opposite direction from the integrated virus as the *ARL15* probe (Figure 5.12a). A representative image is shown in Figure 5.12b. Analysis of the 3D distances (x, y and z plane) indicated that in the integrated chromosome, the HPV16 probe and the *ARL15* probe were significantly closer together than the HPV16 probe and the control probe (Figure 5.12c and d). Additionally, when comparing the distances between the control probe and the *ARL15* probe in both, the integrated and non-integrated chromosomes, the two probes were significantly closer together in the chromosome where HPV16 had integrated (Figure 5.12e and f). This suggests that HPV16 integration affects host genome architecture, resulting in a long-range interaction between *ARL15* and the viral genome.

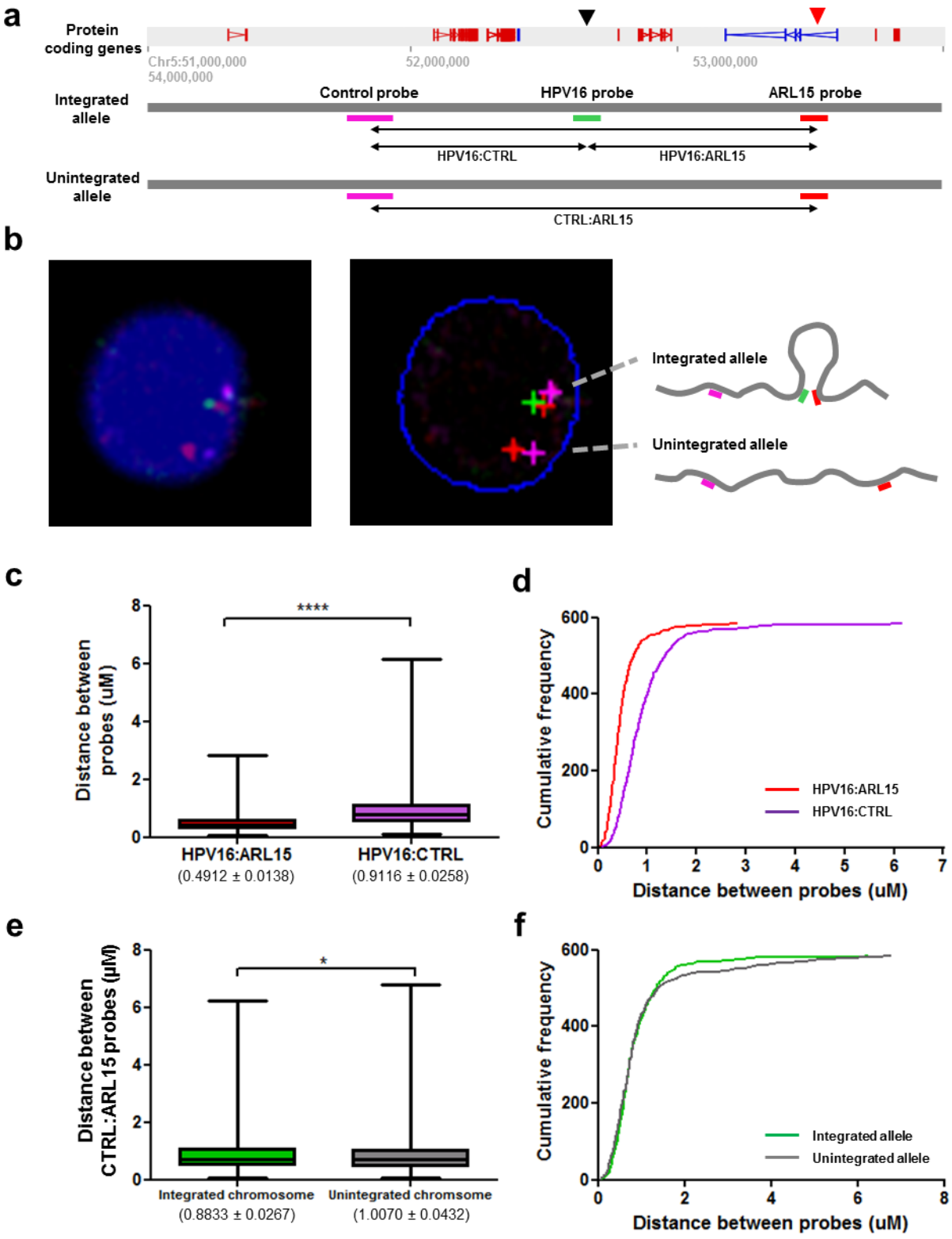


Figure 5.12 | Detection of HPV16-host genome 3D chromatin interaction in clone G2 by fluorescence in-situ hybridisation FISH

(a) Schematic showing the positions of the DNA probes used on the integrated and non-integrated alleles of a portion of chromosome 5 in W12 G2. The control probe (purple) hybridises to a region of the host genome (chr5:51,676020-51873551), distance matching but in the other direction to the *ARL15* probe (chr5:53,473,886-53,584,235) coloured in red. The HPV16 probe is depicted in green. **(b)** Representative image of

the three probes hybridised to W12 G2 genome in a 3D FISH experiment (*left:* raw image with merged channels*; middle:* MetaCyte analysis result*; right:* suggested locus conformation). Analysis of the 3D distance between both sets of 3D FISH probes: HPV16:control (purple) and HPV16:*ARL15* (red), in the copy of chromosome 5 that contained the viral genome is shown in **(c)** a box whisker plot and **(d**) a frequency distribution plot. Analysis of the 3D distances between the "control" and "interacting" probes in both the integrated (green) and the non-integrated (grey) alleles are shown in **(e)** a box whisker plot and a **(f)** frequency distribution graph. n = 585; data presented as mean ± SEM; using unpaired, two-tailed Students T-test: * $p < 0.05$, 888 $p < 0.0001$.

In clone D2, the viral genome had integrated into an intron of the large gene, *Tenm2*, residing on chromosome 5 (Figure 5.13a), far away on the linear sequence compared to the integration site found in G2. Nevertheless, similar to G2, we were able to identify several short- to medium-range interactions. The majority of these interactions were formed with downstream regions of the integrated virus. The interacting loci were all residing within this large host gene, 49 to 527 kb away on the linear sequence, and were all found to be either active open regions or CTCF binding sites (Figure 5.13b).

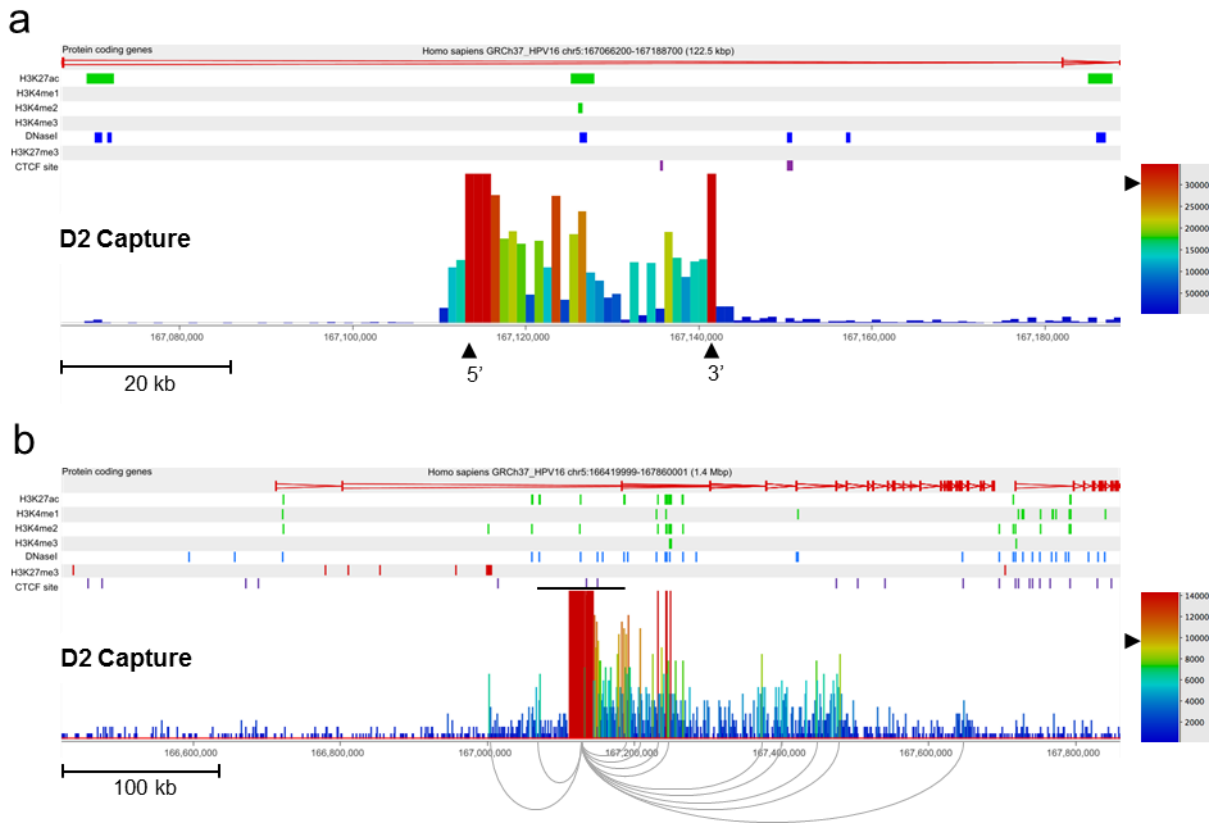


Figure 5.13 | Identification of short and long-range interactions between integrated HPV16 and the host genome in W12 clone D2

(a) SCRiBL data 122.5 kb across the HPV16 integration locus in W12 clone D2. The highest profile peaks are reflecting the integration sites and were labelled with arrowheads. **(b)** SCRiBL data across a 1.4 Mb locus surrounding the viral integration site in W12 clone D2. The black line seen above the locus indicates the genomic region seen in the panel above. Arches are drawn for interactions, with more than 6,000 normalised reads. For both panels, the key indicates the normalised rad counts. Additionally, protein coding genes were colour coded based on their orientation (red = + strand; blue = - strand) and activating histone modifications are shown in green, whereas

repressive H3K27me3 is represented I red; CTCF binding site are indicated in purple and DNaseI sites are shown in blue.

In summary, SCRiBL libraries enabled the detection of short- and long-range interactions between the host and the integrated viral genomes, which were confirmed by microscopy. These contacts seem to be predominantly formed with open and active regions or CTCF binding sites.

5.3.5 HPV16 integration can disrupt local host genome architecture, leading to changes in gene expression of the adjacent genes

To evaluate the nuclear architecture of the host genome and to determine if there are any major changes caused by HPV16 integration, we also sequenced Hi-C libraries in biological replicates from the W12 clones G2 and D2, of which both were the only ones to display long-range interactions in the SCRiBL data. The integration sites of clones G2 and D2 are distinct and, as such, they were used as a control for one another. We generated corrected contact matrices around 5 Mb of each of the integration sites at 50 kb resolution. Around the integration site of G2, multiple self-interacting loci can be seen as “triangles” on the heatmap, all within the size of ~1 Mb, most likely reflecting TADs (Figure 5.14a). When comparing the structure of the non-perturbed locus, from the other clone D2, no obvious differences can be identified. The nuclear architecture of clones G2 and D2 was additionally evaluated at 2.5 Mb either side of the D2 HPV16 integration site (Figure 5.14d and e). As with G2, both clones exhibit similar host architecture across the region and no disruption or generation of self-interacting domains can be observed. Furthermore, when plotting the insulation scores around the integration sites in both clones, with the other clones serving as a control, no differences in the insulation scores was detectable (Figure 5.14c and f). The insulation score is the highest around TAD boundaries and a local minimum relates to the centre of a domain. In both cases, the virus has integrated into the central part of a domain without any alteration in the surrounding domain structure.

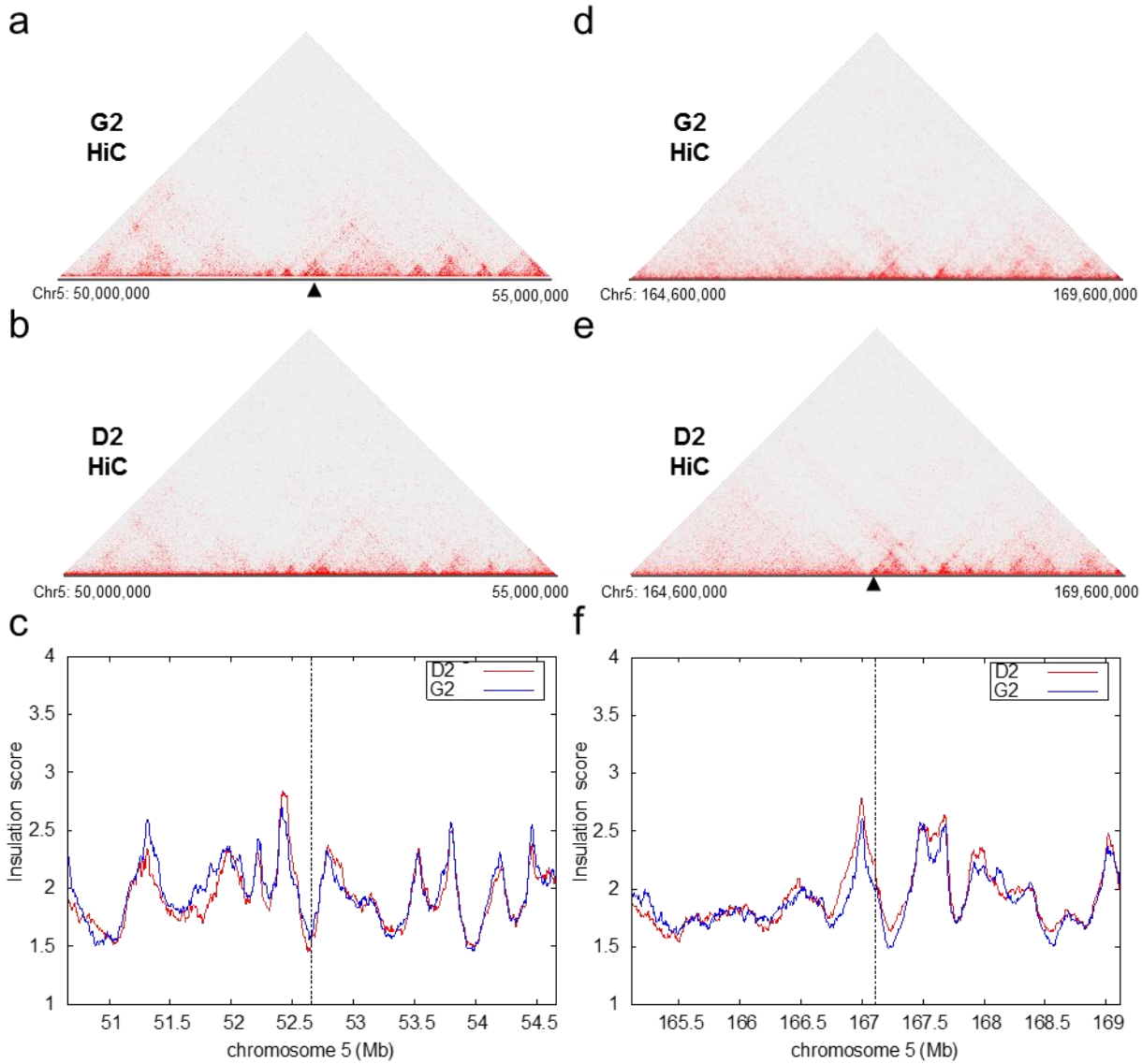


Figure 5.14 | Changes in host genome architecture and domain boundary strength upon HPV16 integration

Corrected heatmaps of 5 Mb across the G2 integration locus (chr5:50 Mb – 55 Mb) for **(a)** clone G2 and **(b)** W12 clone D2. An arrowhead depicts the integrated viral genome. **(c)** Insulation score across the same locus as above for clone G2 (blue) and D2 (red). Corrected contact matrices were also calculated for the 5 Mb across the integration site in clone D2 (chr5:164.6 Mb -169.6 Mb) for **(d)** clone G2 and **(e)** W12 clone D2, with the integration site being depicted by a black arrowhead. (f) Insulation score calculation obtained for the D2 integration locus in clone G2 (blue) and clone D2 (red).

Next, we plotted differential heatmaps around the integration sites and overlaid the obtained matrices with the interaction frequencies obtained from the respective SCRiBL experiments. We further added previously published TAD boundary information (Dixon et al., 2012), which made clear that all interactions originating from the integrated viral genome in clone G2 are falling within the same, unaltered TAD (Figure 5.15b). Strikingly, we could detect a loss of interactions with the rest of the TAD, of the *ARL15* intron, which is now interacting with the viral

genome. The same was observed for a CTCF binding site ~100 kb upstream of the integration site (Figure 5.15a).

Moreover, changes of host gene expression were evaluated by using previously generated RNA-Seq data (Coleman group) for seven W12 integrant clones (A5, B, D2, F, G2, H and R2). Bioinformatic analysis of RNA-Seq data was conducted in collaboration with Anton Enright (EBI-EMBL). The expression of the protein coding host genes 2.5 Mb either side of the HPV16 integration site were compared with the average of the six other control clones, to determine whether host gene expression changed as a result of integration. As shown for G2, changes in both directions, up and down regulation, can be observed, with the gene involved in the long-range contact. *ARL15* being significantly upregulated upon integration close by.

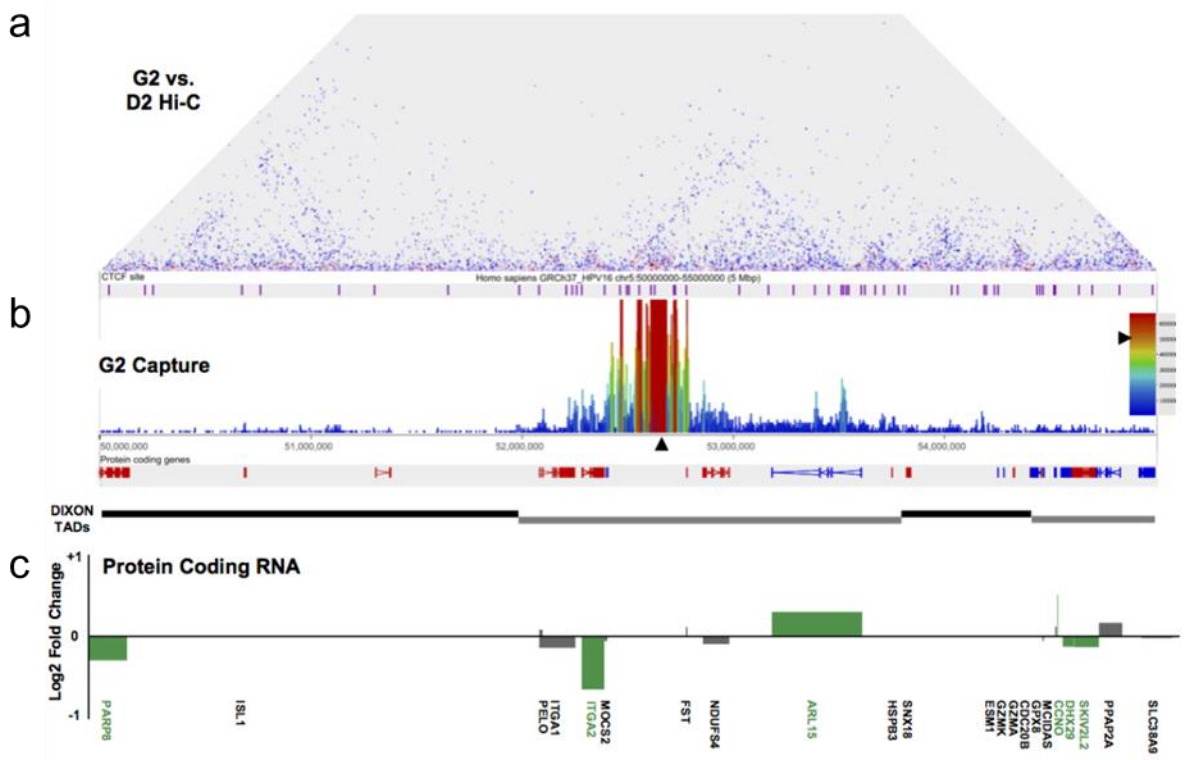


Figure 5.15 | Changes in host genome architecture and gene expression caused by HPV16 integration in W12 clone G2

(a) Differential heatmap of clones G2 and D2 obtained for the G2 integration locus. Blue squares represent host-host interactions, less frequently observed in G2, while red squares represent interactions more frequently observed in G2 Hi-C libraries. **(b)** SCRiBL data showing 3D interactions between the integrated viral genome and the host. CTCF sites are depicted in purple, genes are colour coded based on their orientation (red = forward; blue = reverse strand). TAD boundaries previously published (Dixon et al., 2012) are depicted underneath. **(c)** Chart indicating the relative expression of W12 G2 clone protein-coding genes 2.5 Mb either side of the HPV16 integration site. Depicted are log fold-changes of host gene expression in the clone G2 compared to the other six-clone average. Significant changes are depicted in green, whereas grey colouring indicates non-significant changes ($p < 0.05$, negative binomial Wald test).

Clone D2 showed less pronounced long-range interactions in the SCRiBL data and thus changes in the differential Hi-C heatmap are less far reaching, compared to G (Figure 5.16a). All interactions of the viral genome are formed within the same, unaltered TAD (Figure 5.16b). Viral integration seems to lead to more local interactions of that locus and only the region just upstream of the integration site seems to lose local contacts in both directions. Changes in gene expression within the integration locus were assessed as before and revealed that the gene, the virus has integrated into in D2, is significantly upregulated and so are the three genes in the neighbouring downstream TAD (Figure 5.16c).

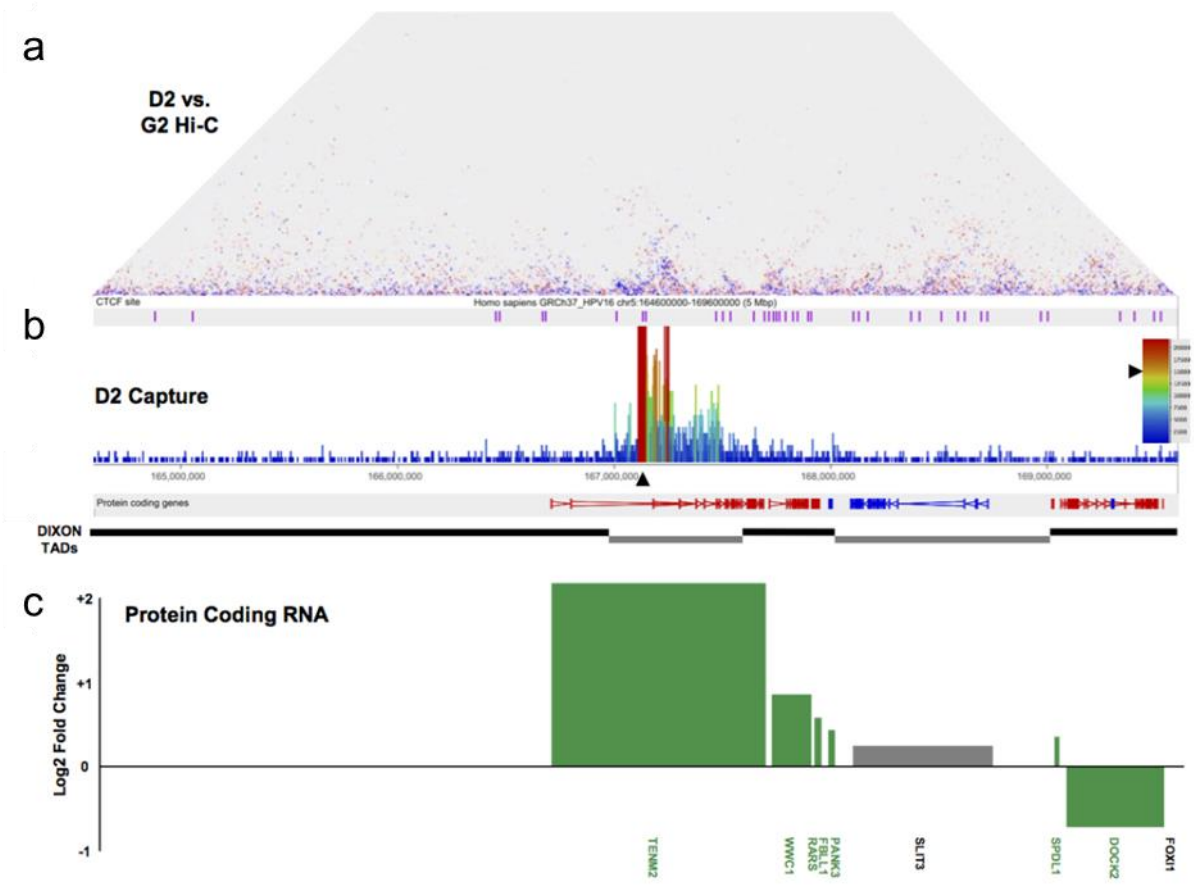


Figure 5.16 | Changes in host genome architecture and gene expression caused by HPV16 integration in W12 clone D2

(a) Comparative Hi-C contact matrix of D2 and G2 across the 5 Mb of the D2 integration locus. Blue squares depict weaker interaction in D2 compared to G2, while red squares represent the opposite. **(b)** SCRiBL interaction between the integrated viral genome and the host genome in clone D2. Genes are colour coded based on their orientation (red = forward; blue = reverse strand) and CTCF sites are shown in purple. Previously published TADs (Dixon et al., 2012) are displayed underneath. **(c)** Log2 fold-changes of relative expression levels observed in clone D2 compared to the other 6-clone average. Green represents significant changes, while grey marks non-significant changes ($p < 0.05$, negative binomial Wald test).

Finally, we wanted to examine the changes in genes expression for all five clones used in this thesis. Therefore, we performed the calculation as done before for G2 and D2. In each of the W12 clones analysed (G2, D2, H, F and A5) significant changes to protein coding host gene

expression in both directions were seen across the entire 5 Mb regions. Most notably, where HPV16 had integrated within a host gene, the expression of that gene was consistently significantly upregulated; the change for *TENM2* expression in clone D2 was 4.79-fold greater than the 6-clone average, *MAPK10* expression was increased by 4.47-fold and *RASSF6* increased by 1.62- and 1.64-fold in clone A5 and F, respectively. In addition, the HPV16-host 3D interaction to the first intron of *ARL15* in clone G2 led to a significant increase ($p<0.05$) in expression of the gene with a 0.30-fold increase compared with the six-clone average (Figure 5.16c).

To further investigate the effect of HPV16 integration on host gene expression, the variance in gene expression in the genomic regions adjacent to the HPV16 integration sites was compared with that of the whole chromosome. Host genes, including both protein-coding and non-coding genes, either side of the HPV16 integration site were grouped into bins, each containing five genes. The range and variance of gene expression of each bin was plotted against the mean level of gene expression across the whole chromosome indicating that across the W12 clones expression of genes in this regions was highly variable (Figure 5.17a). In each of the clones analysed (G2, D2, H, F and A5), the variance in gene expression of multiple bins within genomic regions adjacent to the HPV16 integration site were highly significant ($p<0.05$ and $p<0.001$), indicating that integration of HPV16 has a direct influence on host gene expression at and around the integration site (Figure 5.17b). Notably, significant changes to the variance in host gene expression were felt within bins directly at the site of virus integration (W12 A5) or within bins very close to this region (< 6 bins).

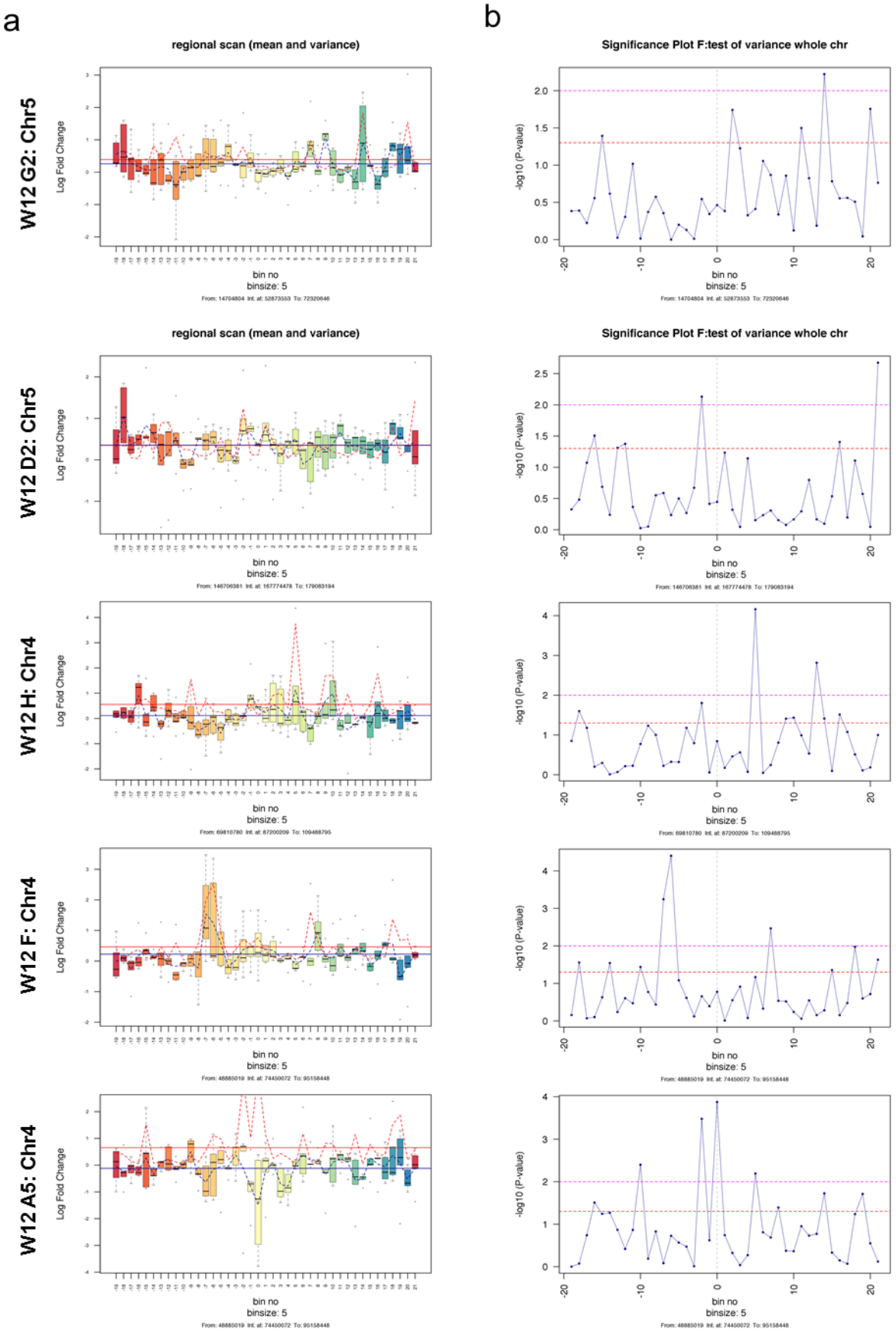


Figure 5.17 | Variance in host gene expression across the host genomic region containing the HPV16 integration site

Each panel in **(a)** indicated the range and variance of host gene expression in the W12 integrant clones as indicated, focusing on the 100 genes either side of the HPV16 integration site. For each clone, gene expression levels were compared to the other six-clone average (based on full dataset: G2, D2, H, F, A5, B and R2). In each panel, the integration site is centred at bin 0 and each bin contains 5 genes. The box whisker plots illustrate the range of expression values within each bin, with bars indicating the median value, the box the interquartile range (IQR) and the whisker the range. The solid blue line indicates the mean gene expression across the whole chromosome, while the dotted blue line shows mean expression levels per bin. The mean variance of gene expression across the whole chromosome is plotted as a solid red line, while a dotted red line depicts the mean variance per bin. Each of the right hand panels **(b)** shows the significance of the variance in gene expression within each bin. Each point represents a five-gene bin, corresponding to those on the left hand side. The horizontal lines indicate the significance of the variance in each bin, compared with variance in gene expression across the whole chromosome (above the dashed red line $p < 0.05$; above the dashed pink line $p < 0,01$).

In summary, we could demonstrate by combining expression data and data containing information about 3D genome architecture, that gene expression around the analysed integration sites is highly variable and significantly altered. This is partly but not exclusively mediated by DNA-DNA loops originating from the integrated viral genome.

5.4 Discussion

Genomic instability and structural abnormalities of the genome are hallmarks of many cancers, some of which are caused by viruses; especially HPV16 is highly associated with cervical carcinomas. Previous studies have shown that HPV16 preferentially integrates into CFS and that integration can lead to induction of gene expression for genes in close vicinity. By capturing the HPV16 genome from Hi-C libraries and genomic DNA libraries, we were able to locate the exact integration sites for a panel of five W12 clones, which represent early stages of HPV16 caused carcinogenesis. Hi-C and SCRiBL revealed that viral integration does not disrupt host genomic domain structure. Instead, we found specific short- and long-range contacts originating from the viral genome, within pre-formed and non-altered TADs. Integrative analysis of RNA-Seq and SCRiBL revealed a striking correlation between viral integration and transcriptional regulation.

In the first part of this chapter, I established and validated a novel methodology to capture small genomic regions of interest from Hi-C and genomic DNA libraries. Both methodologies are of high value for research on small integrating DNA loci, such as integrating viruses like HPV and HIV. First, the resolution of a chromosome conformation capture assay is ultimately determined by the frequency with which the restriction enzyme fragments the genome. In conjunction with the small ~8 kb HPV16 genome, only the 4-cutter MboI resulted in sufficient resolution on the viral genome and had previously been tested in Hi-C experiments (Fudenberg et al., 2016). This additionally resulted in more complex libraries, with complexity relating to the number of theoretically possible interactions within the library. In the case of human Hi-C

libraries generated with MboI, there are ~16 million fragments with the staggering possibility of 100 trillion pairwise interactions, although physical linkage and the very small frequency of interactions in *trans* (Lajoie et al., 2015) will dramatically reduce this number. Given the enormity of theoretical interactions and the small size of the viral genome, capturing the HPV16 genome was necessary. This reduced the overall library complexity and enriched for virus-host interactions and can enable analysis at restriction fragment level (Schoenfelder et al., 2015a). Since the first part of the SCRiBL protocol is a Hi-C library generation, we could assess and confirm high quality of the libraries prior to sequencing by PCR amplification of known short range interactions and a digestion assay. Following the principles of 3C conformation capture, two forward primers ~5 kb away on the linear sequence were designed towards the ends of restriction fragments. The *RPL13A* genomic locus was used as a control region as it had previously been shown to be abundant and stably expressed in the W12 integrant clones (Scarpini et al., 2014). The detection of a band with the expected size suggested successful restriction enzyme digestion and subsequent ligation. The occurrence of multiple other bands on the gel, roughly resembling a ladder with 250 bp spacing, suggests that ligation results in concatemers of multiple fragments ligated to each other. Due to the short length of MboI fragments it is possible to amplify over multiple fragments ligated to each other, resulting in the observed ladder. Thus, we designed a 4-cutter specific control by designing primer pairs that specifically amplify a ligation product between two fragments. This furthermore showed the successful digestion and ligation of then Hi-C libraries. The crucial step that separates 3C libraries from Hi-C libraries is the fill-in with biotinylated oligonucleotides, enabling subsequent enrichment for successful ligation events. To assess the fill-in efficiency, we utilised the same control as for the 6-cutter libraries generated as part of Chapter 4. This revealed overall high efficiency. Of note, as previously mentioned there are 16 million potential interactions per fragment in a library generated using a 4-cutter restriction enzyme. Therefore, assuming 100 % ligation efficiency, it would be possible to detect all ligation interactions starting with just four million cells (4^4 million = 16 million). For the generation of each W12 Hi-C library a starting material of fifteen million cells was used; despite a maximum loss of efficiency of 28.9 % all possible interactions are still recovered ((15 million × 0.711)^4 > 16 million).

The final Hi-C libraries were enriched for the viral genome by in-house custom-made biotinylated RNA oligonucleotides. Capturing Hi-C libraries based on predetermined regions of interest, such as cancer risk loci and gene promoters, has been conducted in an increasing number of studies (Dryden et al., 2014; Jager et al., 2015; Javierre et al., 2016; Schoenfelder et al., 2015a; Schoenfelder et al., 2015b), including chapter 4 of this thesis; however enriching a Hi-C library for an integrated short viral genome is novel. The baits were designed similarly to ones used before (Schoenfelder et al., 2015a), but we needed only 16 to capture the tiny viral genome. We designed a gBlock ® Gene Fragments based approach that allowed for

site-specific T7 promoter ligation, enabling *in vitro* transcription using biotin-UTP. To ensure controlled and specific ligation, T7 promoter adapters were designed with a compatible, cohesive end (BamHI overhang) to that of the 5'-end of DNA fragment (BglII overhang). Ligation of the two DNA molecules generated a new restriction site (5'-GGATCT-3') that could not be cleaved by either enzyme used in the digestion reaction. Following the enrichment step, test PCRs were conducted as usual to determine the optimal number of PCR cycles needed for the final amplification. Furthermore, test PCRs indicated that the post-capture DNA concentration of the HPV16-negative NCx library was lower compared to the W12 integrant clones regardless of equal starting concentrations, illustrating that the biotinylated-RNA baits successfully enriched for DNA fragments containing the HPV16 genome.

The final quality of the library can only be assessed after sequencing and the HiCUP (Wingett et al., 2015) provides useful outputs to do so. Hi-C protocols using 6-cutter restriction enzymes are well established and generally result in high quality libraries with a high percentage of valid reads (see Chapter 4). In a 4-cutter library there are theoretically 16 times more restriction fragments that first need to be cut, then filled in, ligated, pulled down and amplified. Simple upscaling of enzyme concentrations is not always possible, hence suboptimal reaction setups will result. This is reflected by the lower percentage of valid reads obtained from the Hi-C libraries, but especially some of the SCRiBL libraries. The three high valid read number Hi-C libraries result in an even higher percentage of valid reads following the capture step. This has been reported for 6-cutter Hi-C libraries before (Chapter 4). We did not sequence any other Hi-C libraries, but the other SCRiBL libraries are lacking far behind in terms of valid read pairs. Strikingly, the clones where there are three copies of the viral genome per cell, show overall a much better library quality, although this should in theory not affect the quality of Hi-C libraries, as the quality improving step of capturing fragment ends is not performed. The main cause of invalidity in the other SCRiBL libraries and the one Hi-C library are same fragment dangling ends and re-ligation events. Re-ligation events occur when short fragments, such as most of the ones in 4-cutter Hi-C libraries, cannot find other ligation partners than the one they were close to before digestion. Inversely, when fragments are not being ligated, but reside in the library despite the pull-down, they end up as non-informative reads in the sequencing data. This can be caused by inefficient ligation and biotin-removal from non-ligated ends. These steps are candidates for improvement in future experiments.

Hi-C libraries are limited to a resolution of restriction fragments, which was not sufficient to determine viral integration sites with nucleotide resolution. Sequencing of genomic DNA libraries has previously not been successful to determine the integration site, most likely due to the tiny size of the viral genome compared to the host genome (~ 400,000-fold excess). Capturing the viral genome was inevitable, but the SCRiBL baits were targeting only MboI restriction fragment ends, thus missing out on large portions of the HPV16 genome, these baits

were not suitable for gDNA capture experiments. Multiple studies have conducted similar experiments to determine the viral integration sites using tissue samples, at different stages of carcinogenesis, for a range of HPV genotypes (Holmes et al., 2016; Liu et al., 2016; Wang et al., 2013). In each study, extracted DNA was prepared into a sequencing library and enriched for the HPV genome via the hybridisation of HPV genome-specific probes; however, the design and manufacture of HPV probes used was carried out by external companies including MyGenostics Inc. (Liu et al., 2016; Wang et al., 2013) or Roche NimbleGen Inc. (Holmes et al., 2016). Here, we PCR amplified the viral genome in four consecutive, slightly overlapping stretches, with each of the four forward primers containing the T7 promoter sequence, allowing for subsequent *in vitro* transcription using biotin-UTP. Since the primer pairs were slightly overlapping and the *in vitro* transcription was not always complete, resulting in a RNA smear prior to fragmentation, not even coverage of the viral genome by biotinylated RNA baits was achieved and hence this capture system is not quantitative. For enrichment of quantitative experiments, such as ChIP, a new capture system, evenly covering the viral genome would need to be generated.

The methodology presented in this chapter lays the foundation for furthering our understanding of HPV16 virus integration and associated selection of cells in the field of papillomavirus biology. Viral genome integration represents a crucial step in tumorigenesis (Pett et al., 2007) and elucidation of integration events is an essential requirement for understanding HPV-induced carcinogenesis. Coupled with SCRiBL Hi-C analyses, further levels of virus and host genome regulation can be identified. Changes to gene expression because of virus integration and long-range interactions may begin to explain the mechanisms behind the growth advantage of particular cells present across the cells of a polyclonal LSIL.

First, the identification of the 5' and 3' virus-host breakpoints of five W12 clones (F, A5, D2, H and G2) were described at nucleotide resolution. It is important to note that the HPV16 integration sites in four out of five clones identified in this study differ from those that have been previously published (Dall et al., 2008). This nicely illustrates the power of our new NGS based methodology over traditional PCR based approaches, namely Restriction Site PCR and Amplification of Papillomavirus Oncogene Transcripts (APOT). Additionally, and in contrast to previous findings, the results from this study revealed that the HPV16 integration site in W12 clones F and A5 are identical. Although both clones were isolated from the same mixed population of episomal W12 cells (W12Ser2 p12) it is likely that clone A5 is a precursor of clone F. Differences between the two clones arise after just twelve passes of continuous culture, when clone F acquires additional genomic imbalances and shows increased levels of E6 and E7 oncogene expression (unpublished data Cinzia Scarpini/Mark Pett). Furthermore, in four out of five clones analysed, the viral genome had integrated into a host gene, which is

consistent with previous data reporting that integration sites are significantly more likely to be found in host genes than expected (Bodelon et al., 2016). Preferential integration into gene-rich loci is complemented by evidence demonstrating that in cervical squamous cell carcinomas (SCC) HPV16 integrates into regions of open and active chromatin, determined by DNaseI hypersensitivity sites and H3K4me3 marks (Christiansen et al., 2015; Doolittle-Hall et al., 2015). Our findings corroborate the data presented in the previous studies, when overlapping the integration sites with chromatin modifications observed in NHEK. Most commonly found histone modifications were those associated with active enhancers, a transcriptionally competent environment supportive of viral oncogene expression. This illustrates that the HPV16 genome integration into areas of open and active host chromatin is a hallmark of all integration events, regardless of subsequent selection.

Disruption of the *E2* ORF was found in all five clones, by either deletion (clone H) of that locus or by placing it upstream of the viral early promoter during integration (clones G2, F and A5). This is in agreement with the earliest model of HPV integration that promotes oncogenesis by disrupting the *E2* gene. However, the observed integration events promote oncogenesis in a number of different ways including the disruption of cellular genes and their flanking regions and altering their expression (McBride & Warburton, 2017).

Strikingly we found, independent of the number of integrated viral genomes, only one 5` and one 3` virus-host breakpoint in each of the clones. The observation of fewer breakpoints than the viral copy number has previously been reported and led the authors to hypothesise that the discrepancy is due to the amplification of viral integrants and flanking genomic sequences leading to redundant, identical breakpoints (Akagi et al., 2014). By evaluating the positions of the 5' and 3' breakpoints relative to the host genome it was inferred that there are two distinct mechanisms in which HPV16 integrates into the host genome in the W12 clones, which will be discussed elsewhere (Knight et al.; in preparation). Analysis of the nucleotide sequences of both the virus and the host at each breakpoint in the W12 clones adds to the existing body of evidence suggesting that HPV integration likely occurs through microhomology-mediated repair mechanisms (Hatano et al., 2017; Hu et al., 2015; Liu et al., 2016).

The main aim of this chapter was to identify whether viral integration disrupts host nuclear architecture and if there are any short- and long-range contacts formed by the integrated viral genomes. In W12 clones containing more than one copy of the virus genome, both short- and long-range interactions between the integrated virus and regions of the host were identified. Interacting loops were often associated with regions of the host that carried enhancer marks, suggesting that chimeric promoter-enhancer loops are formed as a consequence of the introduction of HPV16 early (p97) and late (p670) promoters. This is most clearly evidenced in clone G2 where, although virus-host interacting reads were found across the entire virus genome, the greatest percentage came from the restriction fragment covering the *E7* ORF,

which contains the p670 promoter. The significance of distal enhancer-promoter contacts via chromatin looping has been demonstrated by the generation of artificially forced loops, which were found sufficient to highly activate gene expression (Deng et al., 2014). Furthermore, the 3D interactions between HPV16 and the host genome were commonly associated with CTCF binding sites in the host. CTCF is instrumental in transcriptional regulation and controlling higher order chromatin structure and can mediate long-range looping (Ong et al., 2014). It has previously been shown that CTCF is recruited to the CTCF binding sites of the HPV genome (Paris et al., 2015). HPV integration therefore results in the insertion of an ectopic CTCF binding site, which is able to form loops with pre-existing host CTCF binding sites, in a similar fashion to that observed upon HTLV-1 integration (Satou et al., 2016). It had been proposed that CTCF sites need to be in convergent orientation in order to form a loop (de Wit et al., 2015; Rao et al., 2014). It would be interesting to see whether this also holds true for ectopically inserted HPV viral CTCF sites. However, in W12 clone G2 all virus-host interactions reside within the same TAD with CTCF binding sites marking the TAD boundaries. The pre-established TAD structure does not seem to be altered and functions as an insulating environment, potentially preventing more distal virus-host interactions. Nevertheless, a long distance loop between the virus and the host gene *ARL15* was detected within the relevant TAD. The interacting loop between the integrated HPV16 genome in clone G2 and the host genome within the first intron on gene *ARL15* co-localises at a significantly higher frequency that random control probes at a similar distance (Jager et al., 2015). This contact was also seen in FISH experiments. Although what you “FISH” is not what you “C”, 3C based methods and DNA-FISH are powerful methods to show the relation between ligation frequencies and physical distance (Fudenberg & Imakaev, 2017; Giorgetti & Heard, 2016).

RNA-Seq data suggests that HPV16 integration alters host cellular gene expression at least 2.5 Mb either side of the integration site. Although the vast majority of 3D interactions were restrained within TADs, changes in gene expression were observed beyond TAD boundaries. This may be the result of secondary effects, such as viral *E6* and *E7* oncogene expression, which has far-reaching consequences and affects many cellular processes, including host cellular gene expression (Zacapala-Gomez et al., 2016). It has been shown that gene activation can occur via genetic alterations that disrupt insulated neighbourhoods (TADs) as a consequence of aberrant activation by enhancers that are normally located outside of the TAD (Hnisz et al., 2016). Therefore, it is possible that changes to the genomic sequence as a result of HPV16 integration, including deletion and focal amplification, could disrupt pre-existing TAD boundaries causing changes to host gene expression. This mechanism can be excluded for the W12 clones where Hi-C data was obtained, as there does not seem to be a difference in TAD boundary location, nor strength. In the case where the HPV16 genome had integrated into a host gene, we consistently observed transcriptional induction of that gene. The

introduction of an additional, viral promoter and its associated regulatory/enhancer region (LCR) could potentially be the cause for this effect. Amplified gene expression may be caused by increased TF-mediated RNAPII recruitment to the extra target sequence i.e. the p97 promoter (Jonkers & Lis, 2015).

Data presented in this chapter also demonstrate that virus-host 3D interactions affect host gene expression; the interaction between integrated HPV16 and *ARL15* in clone G2 resulted in a small but significant increase in *ARL15* expression. It is hypothesised that in this situation the formation of a virus-host interacting region and the introduction of the virus promoter is sufficient to increase the transcription efficiency of host gene *ARL15* as a result of increasing the local concentration of gene promoters (Krijger & de Laat, 2017). These data indicate that influence of HPV integration can be exerted over greater distances by forming larger virus-host interactions than previously described (Adey et al., 2013). Directly correlating HPV16 integration in a particular W12 clone with changes in host gene expression, by comparing the expression of host genes at and around the integration site with random regions of the genome, is an inadequate method of comparison due to natural fluctuations in gene expression due to copy number variations (CNVs) as well as chromatin structure affecting TF binding. To address this issue we analysed the variance of gene expression across the W12 clones and compared this to the chromosome in which the virus had integrated. This analysis indicated that although significant changes were found across the whole region (100 genes either side of the integration site), they were predominantly found at, or close to, the site of HPV integration. These findings are in keeping with the observation that gene expression levels at sites of HPV integration were significantly higher in tumours with HPV integration compared with the expression levels of the same genes across other tumours without integration at that site (Ojesina et al., 2014).

5.5 Conclusion

The work presented in this chapter shows that chromatin conformation capture methods, such as SCRiBL, can be adapted to generate a novel method to elucidate 3D interactions between the integrated HPV16 genome and the host. Based on the same capture principle but by modifying the assay, we could identify viral integration sites with nucleotide resolution, which led to a potential mechanism of integration. Furthermore, the results presented in this chapter provide evidence that modifications to the host genome as a result of HPV integration that are present in advanced SCCs also occur in pre-malignant integrant cells derived in the absence of selective pressure and are therefore characteristic of all HPV integration events. Further work is needed to establish whether long-range virus-host interactions contribute to the growth advantage and selection of particular cells across the mixed population of a polyclonal SIL.

6 General discussion

The 3D organisation of the nucleus affects gene expression and genome replication. During my PhD studies, I applied Hi-C and SCRiBL, a technique based on capturing specific di-tags of interest from Hi-C libraries, to determine genome-wide association profiles, and changes thereof, upon lytic mCMV infection. In parallel, changes in transcription were assayed over time by employing metabolic labelling of newly transcribed RNA at multiple time points of infection. In short, projects presented in this thesis comprise a detailed spatiotemporal description of 3D genome structure and contacts and genome-wide transcription kinetics in mCMV-infected cells. Additionally, I applied genome architecture profiling methods to cells in early stages of HPV16-induced carcinogenesis and identified correlations between chromatin conformation and gene expression. In this chapter, I will discuss the broader context and implications of the work and especially the techniques presented in this thesis.

6.1 3D genome organisation: Cause or Consequence?

The technology behind many molecular biology methods has undergone a radical transformation in the past twenty years. The efficiency and capacity of next generation sequencing (NGS) techniques has rapidly increased, in a trend that seems set to continue. This has also led to major advances in our understanding of genome architecture over the past few years. Prior to NGS, methods such a microscopy (FISH) and PCR based 3C technologies were used to gain insights into the organization and structure of genomes. In only a short period of time, NGS technologies have grown from yielding a few hundred thousand short reads to being able to produce billions of long reads in less time and for less money. This advancement has driven the development of genome-wide functional assays and of computational methods to analyse this new and exciting data.

Studies have undoubtedly shown that there is a strong correlation between genome conformation and gene expression (Bonev et al., 2016; Krijger et al., 2016). What remains unknown is the causality of events: did the conformational change bring forth the changed expression patterns, or does RNA transcription change the conformation of the loci involved? In order to answer this question, one could imagine performing multiple Hi-C experiments in a time series for major cellular events (such as cell differentiation or stress response), along with a knockout of DNA elements responsible for certain DNA conformations using the CRISPR/CAS9 system (Perez-Pinera et al., 2013). In Chapter 3 and 4, I tried to obtain first indications on the sequence of events of regulated gene expression by assessing host nuclear architecture at different time points of infection. By performing both, Hi-C/SCRiBL and 4sU-Seq, at multiple time points upon infection I was able to shed light on the temporal properties of the nuclear changes involved. Despite the dramatic transcriptional changes, observed as early as 2 hpi (Figure 3.4), and the striking disruption of nuclear architecture

visible by microscopy (Gibbs et al., 2013), I found only minor changes in the overall folding pattern of the host genome, as measured by Hi-C (Figure 4.5). High-resolution contact maps of the *Nfkbia* locus revealed that, even for genes with very large transcriptional changes upon infection, corresponding changes in DNA looping did not occur (Figure 4.11c). This suggests that, at least for mCMV infection, “re-colouring” (where regions interacting with TSS change their activity) is a more prominent feature of fast transcriptional responses than “re-wiring” (where interacting regions of a TSS differ between two samples) (Freire-Pritchett et al., 2017). This does not allow answering the initial question though, if genome architecture is cause or consequence of transcription. It is possible, however, to inhibit transcription genome-wide by using for example α-amanitin to inhibit RNA Polymerase II/III, or Actinomycin D or CDK*9* inhibitor Flavopiridol to inhibit RNA Pol I (Bensaude, 2011). The effects on nuclear structure can be profound. Flavopiridol causes the nucleolus to disintegrate, and triggers a widespread re-localisation of several proteins and RNA species, such as the splicosomal complex or the small nucleolar ribonucleoproteins to form the so-called Dark Nucleolar Caps (DNCs) (Burger et al., 2010; Shav-Tal et al., 2005). The treatment of mouse erythroid cells with α-amanitin had only little effects on DNA-DNA contacts as measured by 3C and 4C (Palstra et al., 2008). From that, it seems that the large-scale chromatin structure is not strongly dependent on transcription. On the other hand, transcription-dependent nuclear structures, such as the nucleolus (and possibly transcription factories), do require transcriptional activity to hold together.

Nevertheless, we can say that, for both genes and regulatory elements (spatial) localisation truly matters (Krijger et al., 2016). Locally, looping interactions are one of the most important ways to modulate gene activity, through enhancer/silencer elements or co-localisation with transcription factories. On a larger scale topological domains are unique in their ability to regulate large sections (hundreds of kilobases or even megabases) of chromatin in unison (Dixon et al., 2012), which expands even further into megabase sized A and B compartments (Lieberman-Aiden et al., 2009).

6.2 TADs: the building blocks of the genome

Hi-C studies have revealed that within their territories, chromosomes are partitioned into large compartments at the multi-Mb scale, with genomic regions either assigned to the active and open A compartment or the inactive and closed chromatin B compartments (Lieberman-Aiden et al., 2009). Changes in the compartments between two samples can uncover large-scale differences between the expression states of the samples. One can use this comparison to pinpoint regions that may show substantial differential gene expression between the genes contained within the regions changing compartment, as shown for some developmental genes upon mCMV infection (Figure 4.9). These tissue specific domains are formed by largely tissue-invariant TADs, which have been shown to be associated with gene-regulatory features and it

is hypothesized that TADs specify elementary regulatory micro-environments in which promoters interact with enhancers (Dixon et al., 2015; Dixon et al., 2012; Dowen et al., 2014). The block-like structure of TADs clearly indicates elevated interaction frequency within a TAD. However, given that we measure a population average and the observed intricate hierarchies of such structures (Rao et al., 2014; Zhan et al., 2017), interpretation of TADs is not straightforward. It has been proposed that TAD-like structure may be driven at least in part by looping interactions between loci located within them (Giorgetti et al., 2014) or by supercoiled DNA structures (plectonemes) (Benedetti et al., 2014; Le et al., 2013). As shown for the *Nfkbia* gene, differential looping between enhancer and promoter elements is not a common feature upon mCMV infection, which might explain why TAD structures do not change upon infection. Additionally, CTCF and cohesin binding have been shown to be enriched at TAD boundaries (Dixon et al., 2012; Van Bortle et al., 2014). HPV possesses a CTCF binding site, which led to the speculation that HPV integration can alter genome organisation on a larger scale. We did not find any evidence for this in our data. Recently, it has been shown that upon CTCF depletion, TAD boundaries become less insulating and loops are lost genome-wide. Additionally, the authors report a role for cohesin in mediating long range interactions and TAD formation, but TADs were not entirely absent upon cohesin depletion, while A/B compartments become more apparent (Wutz et al., 2017). These results indicate an important role of CTCF and cohesin in forming TAD structures, but point towards additional players in TAD formation and maintenance. Since first reported in 2012, tremendous effort has been put in to understand TAD formation and function but to date it remains unclear what physical structures TADs exactly represent and how they are specified in the genome. It has also become clear that the specific structure of a chromosome is highly correlated with the functional output of that chromosome. However, this connection is only a correlation, it has not yet been demonstrated whether TAD structure can cause function or function can create TAD structure. Experiments to further elucidate this relationship by manipulating the genomic elements that control and define a TAD and or experimenting with ways to control or shut down expression, will provide insights into the relationship between structure and function. TADs act as local insulators, limiting a gene to only the enhancers that are contained within the gene's TAD. If this process is tightly controlled, this could have the net effect of lowering a genes total expression. If a gene would sample all enhancers within the genome, tight regulation of that gene may be distributed. Having a tightly regulated micro neighbourhood of enhancers for each gene could represent a mechanism to tightly control gene expression, but also integrated viral genomes.

6.3 Future directions

I predict that in the future, genome structure experiments will become just as common as RNA-Seq or Chip-Seq is today. In fact, given how rapidly new methods such as ATAC-Seq (Buenrostro et al., 2013) have grown in popularity, I suspect that this may happen sooner than

most would think. Using ATAC-Seq as an example, ATAC-Seq signal contains protein/TF specific footprint patterns. This means that given enough depth, from a single ATAC-Seq experiment, one can detect all accessible regions of the genome, and from the footprint patterns, one can infer physical binding of proteins/TFs. Assays that can infer multiple layers of information from a single experiment will be the future. Hi-C can serve as such an assay. Possible applications are: ordering contigs, scaffolding (Korbel & Lee, 2013), detecting translocations or breakpoints, measuring copy number variations or structural variants (Harewood et al., 2017), detecting CTs, measuring genome wide active and inactive compartments (Lieberman-Aiden et al., 2009), detecting sets of nested TAD structures (Weinreb & Raphael, 2016), detecting co-expressed clusters of genes or transcription factories (Osborne et al., 2004), characterizing gene-enhancer looping interactions (Schoenfelder et al., 2015a) and so on and so forth. However, the genome structure field is still in its infancy and the amount of data that can be extracted or inferred will continue to grow in the future. Hi-C may become the method of choice for assembling cancer patient genomes (Harewood et al., 2017). One could also envision using Hi-C data to infer gene expression. Given adequate high-resolution (< 1 kb), one could imagine that expressed genes may have a unique topology or structure compared to inactive genes. From this observed structural difference, one could infer expression status. The same could hold true for protein/TF binding. Specific proteins or TFs could introduce a unique local topology or organization relative to other TFs and alter the neighbouring DNA upon binding. If so – this information could ultimately be used to infer TF association.

Given the advancement and availability of longer reads, up to 60 kb via PacBio Sciences or even up to 100 kb via virtual long read technologies such as those offered by 10X Genomics or Illumina's Moleculo technology, one could envision a Hi-C variant which aims to capture multiple interactions in a single molecule. From this molecule, one could then detect a set of hundreds of DNA fragments that all co-localized in a single cell. This added layer of information could be used to detect mutually exclusive or co-occurring interactions, both of which are currently masked given Hi-C's population average data.

A number of yet unexplored avenues exist within comparative and dynamic nuclear organisation. Comparative studies may yield new understanding of the differences between different tissues, healthy and disease states, the evolution of genome structure and heterogeneity between single cells and populations. Studying the dynamics of genome organisation can give us new insights into cell cycle progression, tissue differentiation, the processes driving nuclear organisation and the effect of pharmacological agents. As our understanding of the gross rules governing these processes increases, we will be able to better understand how differences in nuclear organisation can affect biological function, and how biological function can affect organisation. It is reasonable to expect diagnostic tests based on

nuclear organisations to reach the clinic in the future, along with drugs able to modify organisation, especially as preventative measures.

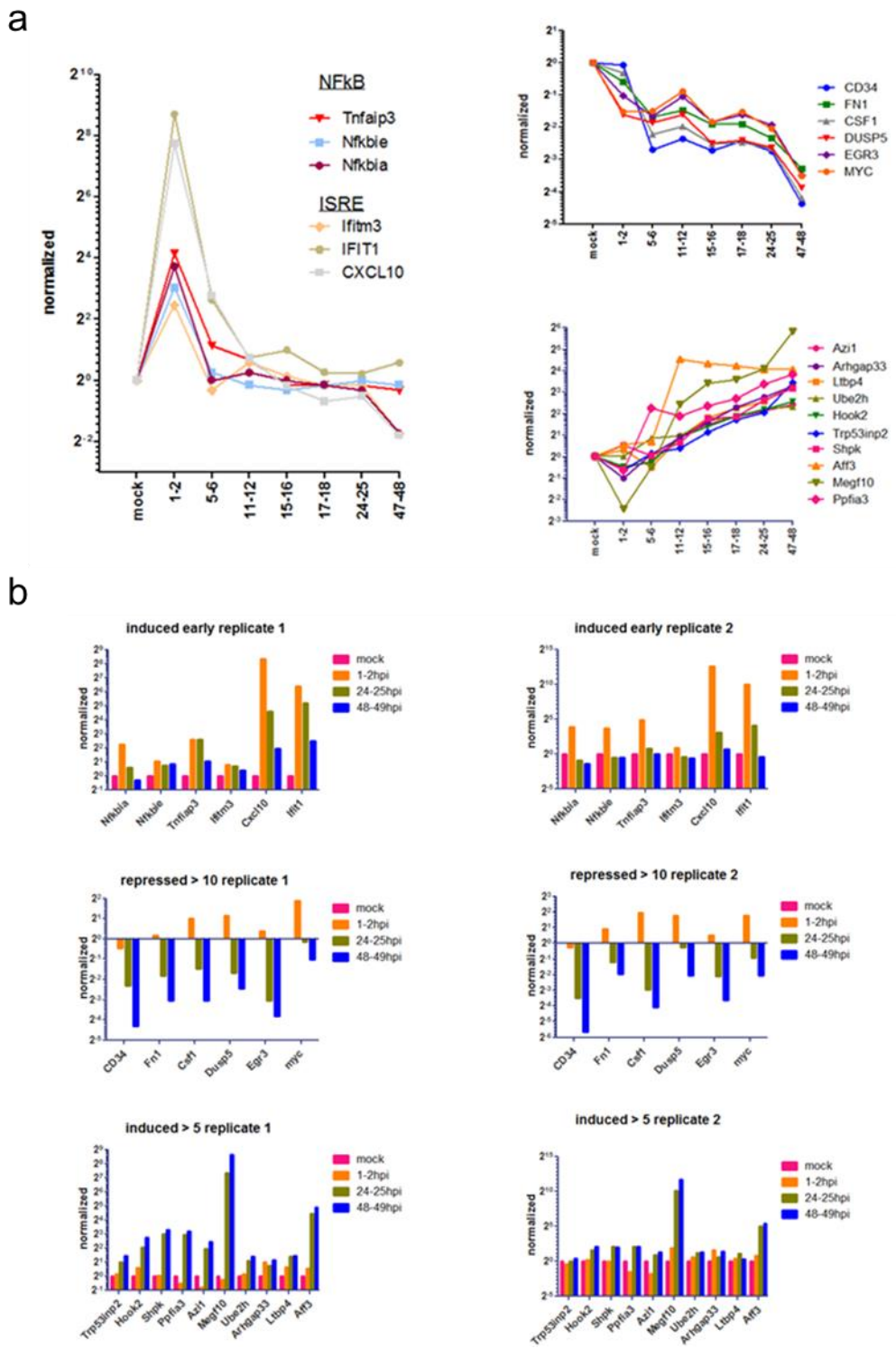
6.4 Concluding remarks

In conclusion, this thesis has made the case for the usefulness of studying the genome structure, and changes thereof, in the context of viral infections. In the case of mCMV infection, a surprisingly stable folding pattern of the host genome was observed, with only A/T-rich lamina associated DNA domains being subjected to compaction. This is despite the observed dramatic changes in transcriptional activity and the previously reported striking genomic rearrangement reported by microscopy studies (Gibbs et al., 2013). Furthermore, by applying novel capture Hi-C technologies to cells in early stages of HPV 16 induced carcinogenesis, I illustrate that many of the host genome alterations seen in advanced squamous cell carcinomas are characteristic of all HPV 16 integration events, independent of selection events later on.

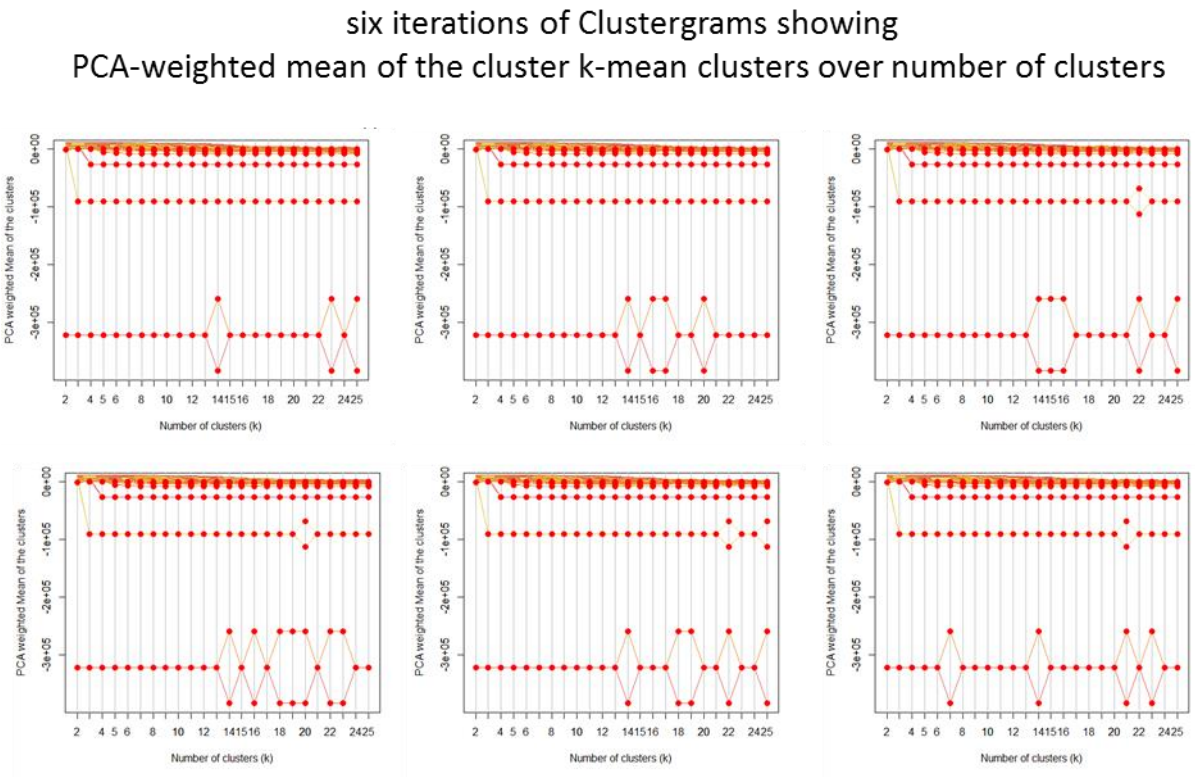
Taken together, the insights gained from the work described in this thesis have made a significant contribution to the complex relationship between genome structure and genome function, especially in the case of viral infection and virus induced carcinogenesis.

Appendices

Supplementary figures

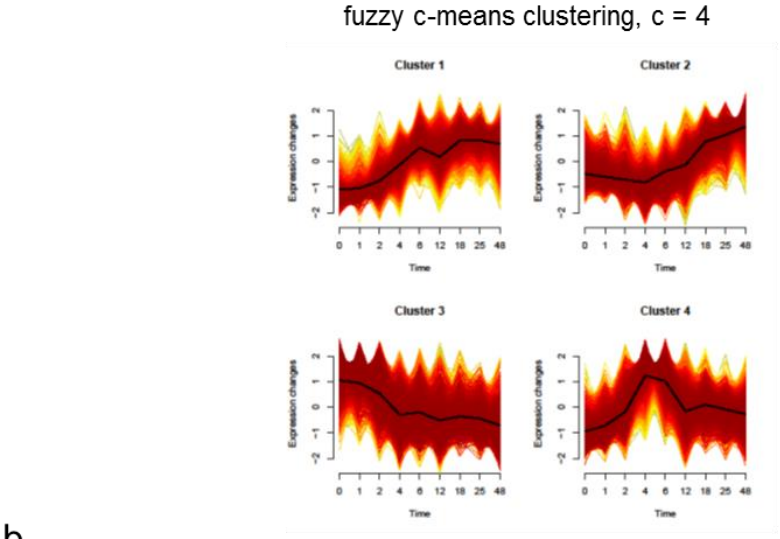


Supplementary Figure 1 | Nascent transcription of genes of interest. (a) Genes of interest were defined based on expression level changes upon lytic mCMV infection observed in previously published data (Marcinowski et al., 2012) and plotted with Graph Pad Prism. **(b)** Real time qPCR performed on cDNA of newly transcribed RNA confirmed the expected transcriptional regulation of candidate genes in both biological replicates.

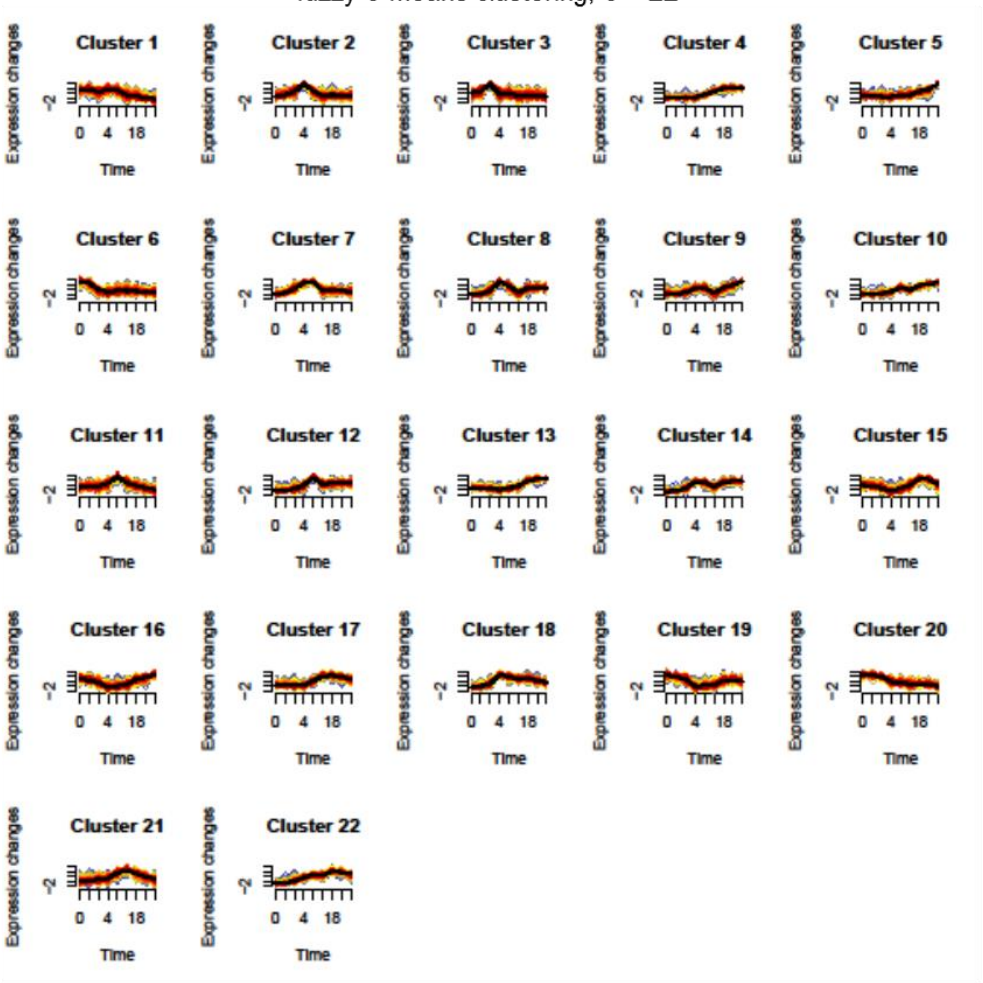


Supplementary Figure 2 | Reproducibility of the cluster number estimated by the Clustergram function in R. The clustergram R package produces a weighted mean of the cluster centres first principal component from each of the iterations. Data points are then ordered according to their clusters first principal component and plotted against the number of clusters. A number of 14 clusters was reproducibly found to be representative of the data over 7 iterations (six shown here and Figure 3.3b)

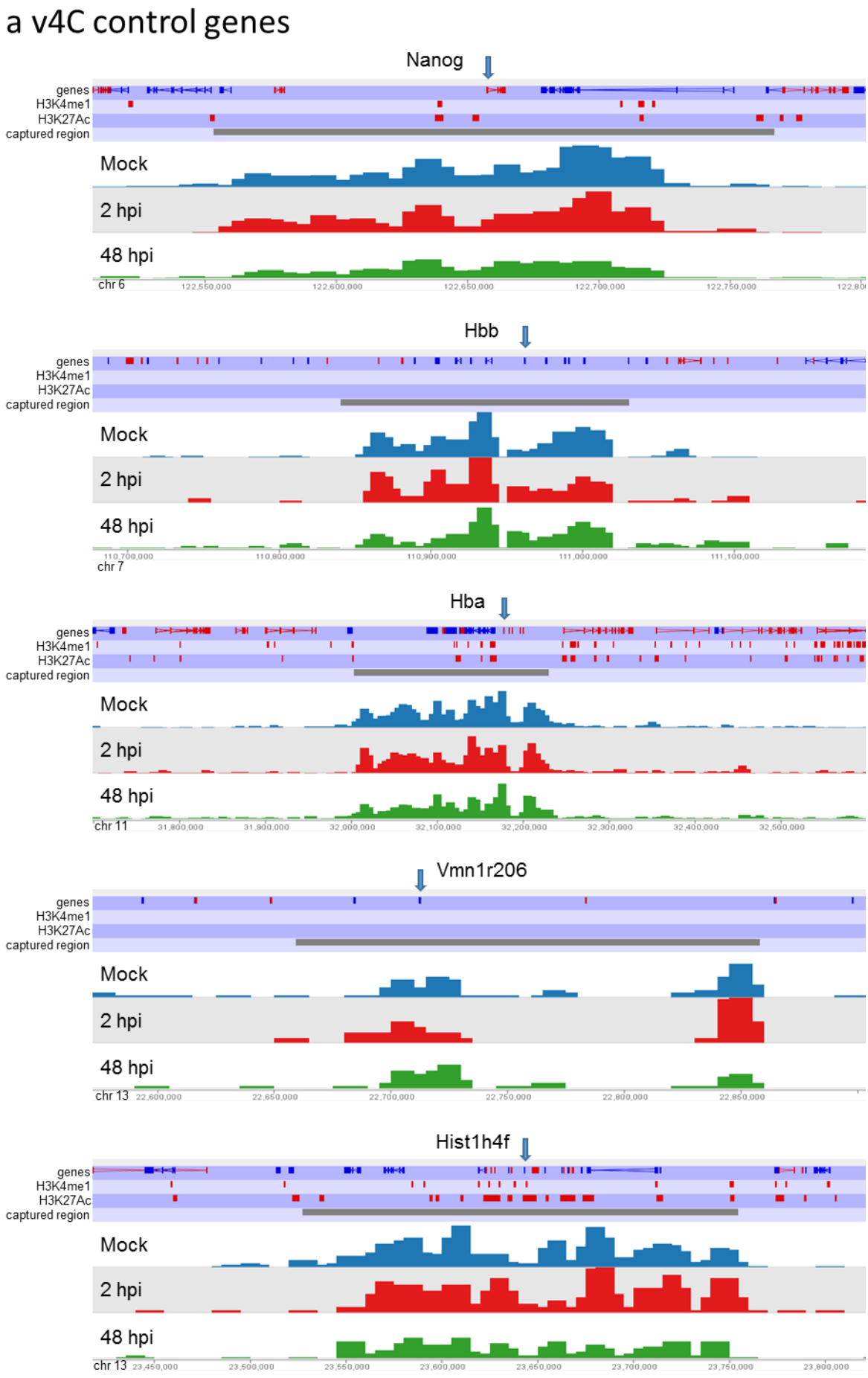
a

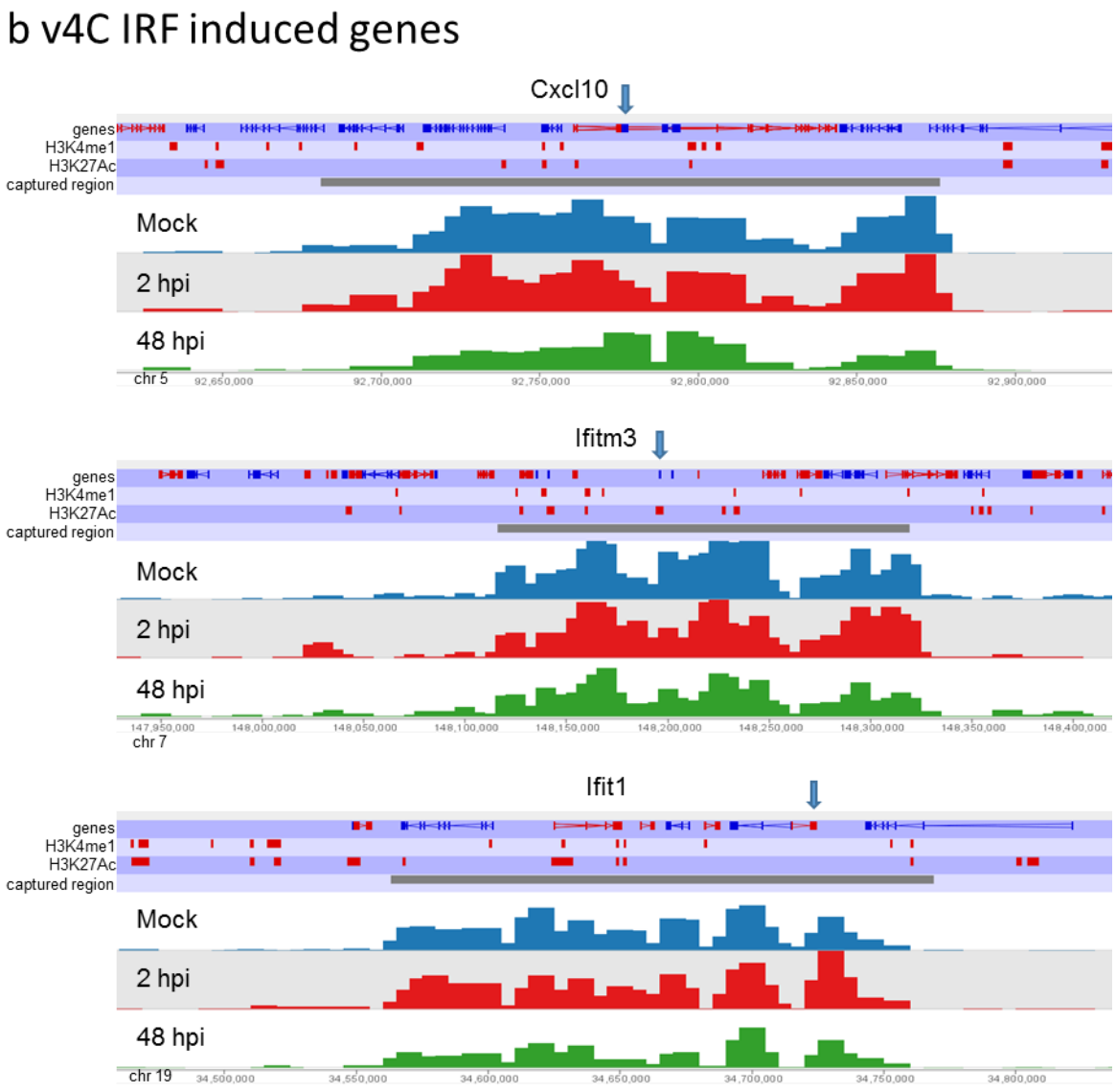


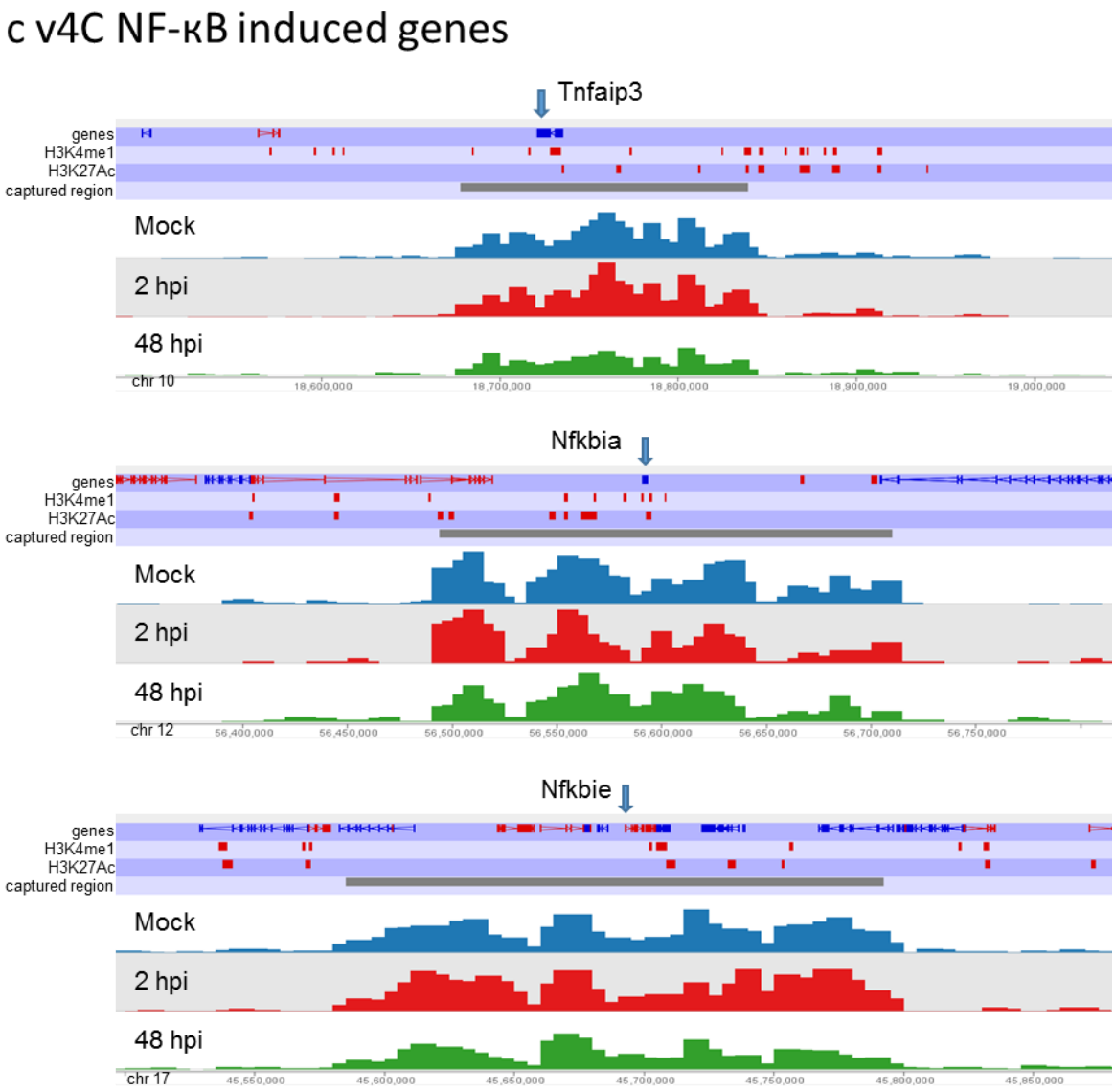
b

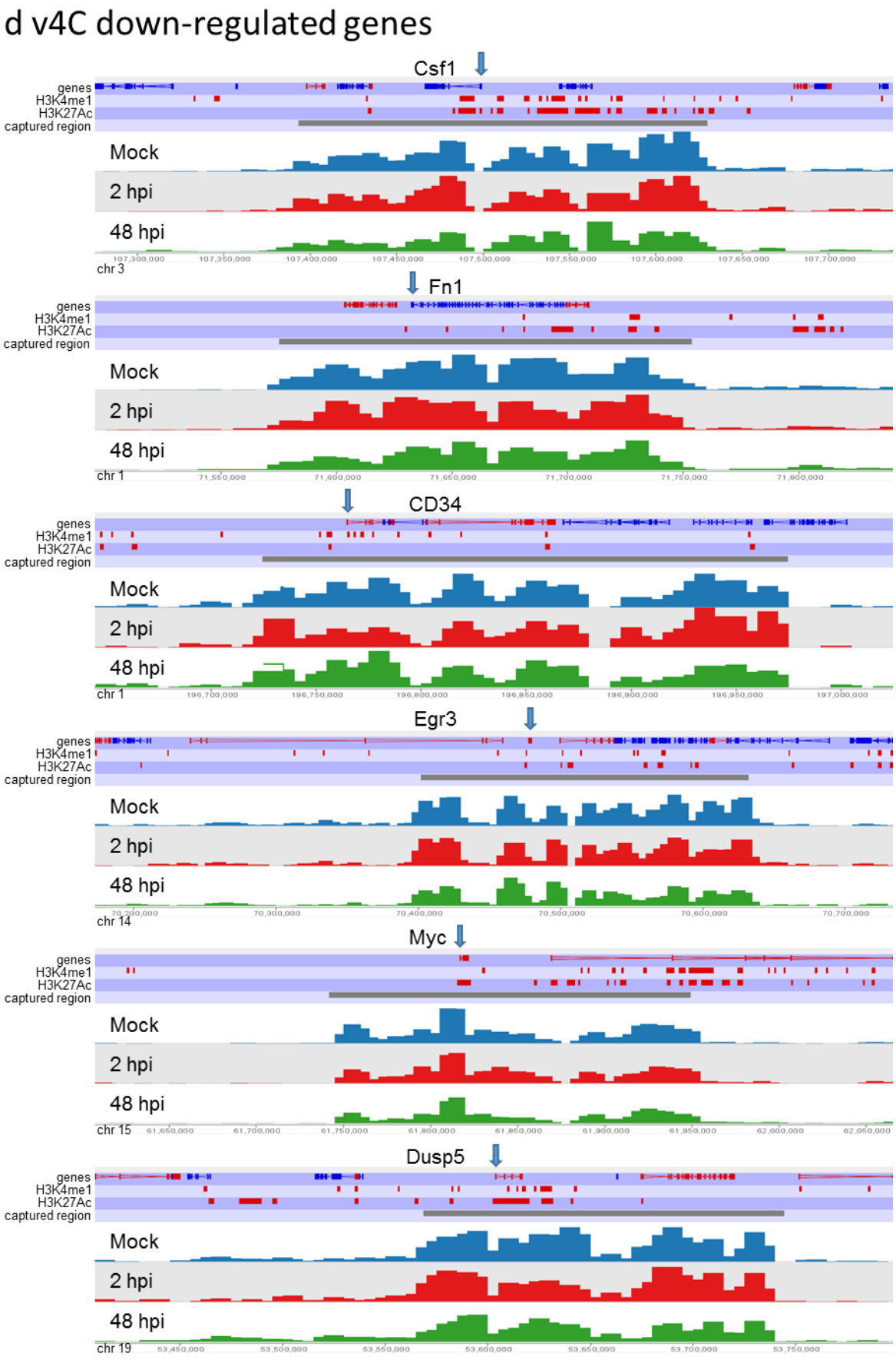


Supplementary figure 3 | Fuzzy c-means clustering results. Soft clustering, conducted with the Mfuzz packages, of averaged RPKM values for transcripts that possessed greater than 0.7 RPKM values in at least one of the time points in at least one of the replicates with **(a)** c= 4 and **(b)** c = 22 clusters. Blue and yellow coloured lines correspond to genes with low cluster membership values; orange and red coloured lines correspond to genes with high cluster membership values. Time point 0 correspond to non-infected NIH-3T3.

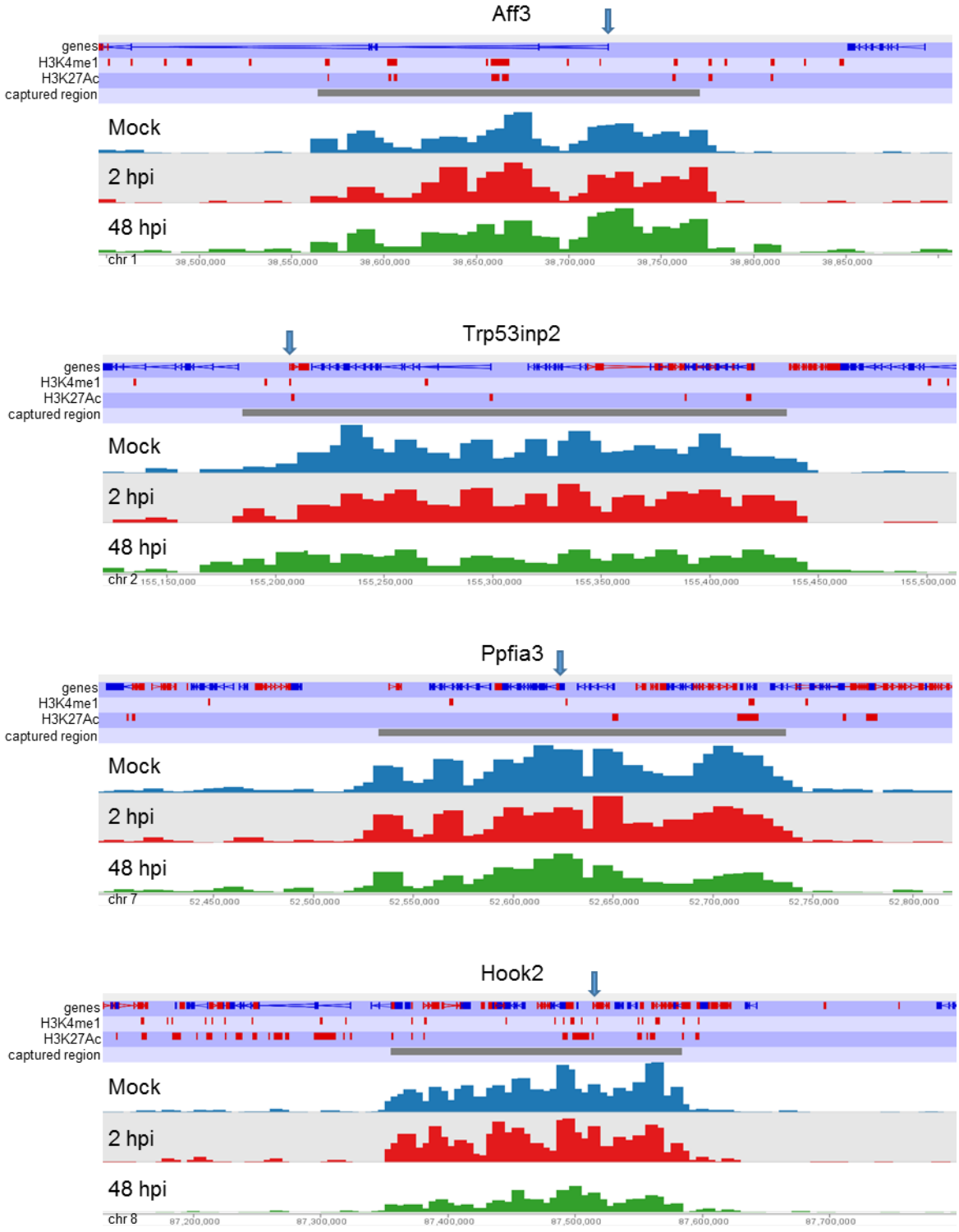

a v4C control genes
Nanog
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 6
122,550,000
122,600,000
122,650,000
122,700,000
122,750,000
Hbb
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 7
110,700,000
110,800,000
110,900,000
111,000,000
111,100,000
Hba
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 11
31,800,000
31,900,000
32,000,000
32,100,000
32,200,000
32,300,000
32,400,000
32,500,000
Vmn1r206
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 13
22,600,000
22,650,000
22,700,000
22,750,000
22,800,000
22,850,000
Hist1h4f
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 13
23,450,000
23,500,000
23,550,000
23,600,000
23,650,000
23,700,000
23,750,000
23,800,000

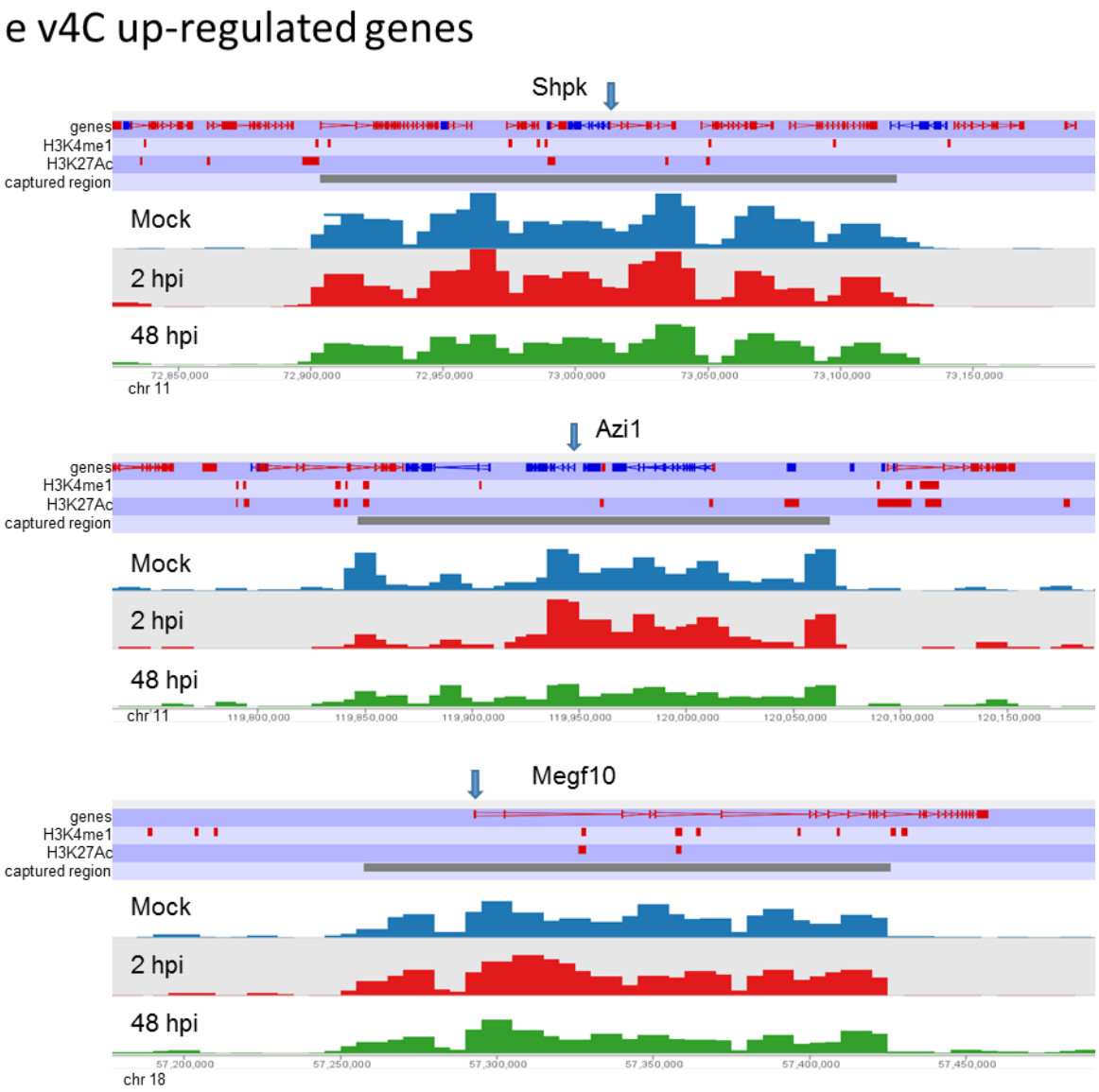

b v4C IRF induced genes
Cxcl10
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 5
92,650,000
92,700,000
92,750,000
92,800,000
92,850,000
92,900,000
Ifitm3
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
147,950,000
148,000,000
148,050,000
148,100,000
148,150,000
148,200,000
148,250,000
148,300,000
148,350,000
148,400,000
chr 7
Ifit1
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 19
34,500,000
34,550,000
34,600,000
34,650,000
34,700,000
34,750,000
34,800,000


c v4C NF-κB induced genes
Tnfaip3
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 10
18,600,000
18,700,000
18,800,000
18,900,000
19,000,000
Nfkbia
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 12
56,400,000
56,450,000
56,500,000
56,550,000
56,600,000
56,650,000
56,700,000
56,750,000
Nfkbie
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 17
45,550,000
45,600,000
45,650,000
45,700,000
45,750,000
45,800,000
45,850,000


d v4C down-regulated genes
Csf1
genes
H3K4me1
H3K27Ac
captured region
Mock
2 hpi
48 hpi
chr 3
Fn1
chr 1
CD34
chr 1
Egr3
chr 14
Myc
chr 15
Dusp5
chr 19

e v4C up-regulated genes





Supplementary Figure 4 | Virtual 4C (v4C) profiles TSS of genes of interest. Genome browser shots of the regions around the indicated genes, showing virtual 4C data from combined replicates with 10 kb resolution; bins were shifted by 5 kb. All data were normalised to sequencing depth, thus enabling direct comparison between the three time points within a given data set (blue= mock, red = 2hpi, green = 48 hpi). Depicted are v4C profiles of genes classified as control genes **(a)**, IF induced genes **(b)**, NF-κB regulated genes **(c)**, down regulated genes **(d)** or genes being up-regulated **(e)**.

Supplementary tables

Supplementary table 1 | HiCUP output statistics for Hi-C and SCRiBL libraries obtained from mouse samples.

	Hi-C Mock Replicate II		Hi-C 2 hpi Replicate II		Hi-C 48 hpi Replicate II		BAC Mock Replicate I	
	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2
total reads	61,145,621	61,145,621	114,942,694	114,942,694	84,250,791	84,250,791	102,820,658	102,820,658
reads after trimming	60,111,431	60,159,672	113,876,649	113,915,919	83,352,592	83,402,549	101,012,119	101,001,448
unique alignments	47,322,285	45,955,536	90,463,321	87,639,835	68,166,802	66,250,946	63,123,399	61,792,000
paired reads	36,107,533		74,086,405		57,136,965		38,058,627	
valid pairs	29,018,507		29,194,988		29,586,956		33,300,286	
invalid pairs	7,089,026		44,891,417		902,974		4,758,341	
same circularised	153,123		1,555,134		4,425,453		638,582	
same fragment dangling ends	727,862		7,223,409		21,997,803		382,793	
same fragment internal	4,806,883		33,723,679		1,001,003		2,040,659	
re-ligation	397,430		971,064		70,289		374,979	
contiguous sequence	18,816		32,217		1,189,434		15,716	
wrong size	984,912		1,385,914		27,550,009		1,305,612	
unique read pairs	28,781,808		28,805,227		27,168,022		31,725,206	
cis close	314,158		602,741		1,816,811		331,825	
cis far	11,697,645		8,806,834		12,646,482		10,937,516	
trans	16,770,005		19,395,652		12,704,729		20,455,865	
on-target	n.a.		n.a.		n.a.		3,527,096	
capture efficiency	n.a.		n.a.		n.a.		11.12	
fold enrichment over Hi-C	n.a.		n.a.		n.a.		203	

	BAC Mock Replicate II		BAC 2 hpi Replicate I		BAC 2 hpi Replicate II		mCMV 2 hpi Replicate I	
	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2
total reads	23,902,291	23,902,291	44,752,343	44,752,343	18,501,405	18,501,405	48,473,570	48,473,570
reads after trimming	23,499,356	23,491,173	44,005,278	44,001,166	18,290,660	18,296,553	47,873,268	47,885,351
unique alignments	14,455,692	14,004,713	27,215,086	26,635,827	10,985,804	10,613,710	38,902,476	38,184,634
paired reads	10,009,964		16,556,970		6,534,770		38,902,476	
valid pairs	1,438,351		12,926,511		1,970,520		10,774,123	
invalid pairs	201,987		3,630,459		249,924		20,572,078	
same circularised	399,102		398,142		463,808		1,460,073	
same fragment dangling ends	709,344		389,101		800,222		515,190	
same fragment internal	95,889		2,013,286		129,384		14,068,354	
re-ligation	4,929		211,234		5,783		1,718,216	
contiguous sequence	27,100		7,616		321,399		83,397	
wrong size	8,571,613		611,080		4,564,250		2,726,848	
unique read pairs	6,904,405		12,337,660		4,481,919		9,815,715	
cis close	90,954		129,737		94,539		122,009	
cis far	2,642,324		4,209,809		1,316,188		3,198,712	
trans	4,171,127		7,998,114		3,071,192		6,494,994	
on-target	1,304,242		1,576,753		814,365		621,335	
capture efficiency	18.89		12.78		18.17		6.33	
fold enrichment over Hi-C	205		212		225		2,561	

	mCMV 2 hpi Replicate II		BAC 48 hpi Replicate I		BAC 48 hpi Replicate II	
	Read 1	Read 2	Read 1	Read 2	Read 1	Read 2
total reads	29,476,879	29,476,879	117,668,797	117,668,797	41,673,873	41,673,873
reads after trimming	29,429,115	29,430,536	115,600,049	115,660,654	41,163,942	41,178,390
unique alignments	28,399,424	27,506,835	84,376,450	83,806,384	27,819,205	27,080,868
paired reads	26,711,920		61,829,454		19,180,520	
valid pairs	24,992,840		24,133,674		5,387,200	
invalid pairs	1,574,207		37,695,780		456,897	
same circularised	9,765,690		1,411,069		1,426,507	
same fragment dangling ends	12,407,533		315,879		2,234,904	
same fragment internal	906,215		5,189,436		466,636	
re-ligation	21,999		1,535,775		40,903	
contiguous sequence	317,196		40,010		761,353	
wrong size	1,719,080		29,203,611		13,793,320	
unique read pairs	1,652,378		23,123,606		13,469,437	
cis close	126,945		771,200		1,014,751	
cis far	638,005		10,109,375		6,413,101	
trans	887,428		12,243,031		6,041,585	
on-target	808,674		6,865,921		3,984,259	
capture efficiency	48.94		29.69		29.58	
fold enrichment over Hi-C	1,516		225		238	

gBlock sequences

AGATCT = BglII

ACTAGT = SpeI

NNN = spacer

AAGCTT = HindIII

Block 1 (1191 nt)

CGTAGATCTATCAAGAACACGTAGAGAAACCCAGCTGTAATCATGCATGGAGATACACCTACATTGCATGAATAT
ATGTTAGATTTGCAACCAGAGACAACTAAGCTTGGCAGATCTTCTCTACTGTTATGAGCAATTAAATGACAGCTC
AGAGGAGGAGGATGAAATAGATGGTCCAGCTGGACAAGCAGAACCGGACAGAGCCCATTACAATATTGTAACCTT
TTGTTGCAAGTGAAGCTTGGCAGATCTCAGCCATGGTAGATTATGGTTTCTGAGAACAGATGGGGCACACAATTC
CTAGTGTGCCCATTAACAGGTCTTCCAAAGTACGAATGTCTACGTGTGTGCTTTGTACGCACAACCGAAGCGAAG
CTTGGCAGATCTTCCTGCAGGTACCAATGGGGAAGAGGGTACGGGATGTAATGGATGGTTTTATGTAGAGGCTGT
AGTGGAAAAAAAAACAGGGGATGCTATATCAGATGACGAGAACGAAAATGACAGTGAAAGCTTGGCAGATCTCTT
GGTCGCTGGATAGTCGTCTGTGTTTCTTCGGTGCCCAAGGCGACGGCTTTGGTATGGGTCGCGGCGGAGTGGTTG
GCCAAGTGCTGCCTAATAATTTCAGGAGAGGATACTTCGTTGAAGCTTGGCAGATCTTCAGAGCCAGACACCGGA
AACCCCTGCCACACCACTAAGTTGTTGCACAGAGACTCAGTGGACAGTGCTCCAATCCTCACTGCATTTAACAGC
TCACACAAAGGACGGATTAACTGTAATAAGCTTGGCAGATCTCAGCAATAGTTTTGCCTTCAACCTTAGGTATAA
TGTCAGGTGGACATGTACCTGCCTGTTTGCATGTTTTATAAAGTTGGGTAGCCGATGCACGTTTTGTGCGTTTTG
CAGAACGTTTGTAAGCTTGGCAGATCTTCAAATATTACAATATGGAAGTATGGGTGTATTTTTTGGTGGGTTAGG
AATTGGAACAGGGTCGGGTACAGGCGGACGCACTGGGTATATTCCATTGGGAACAAGGCCTCCCACAGCTACAAG
CTTGGCAGATCTTCCTTCTATAGTTTCTTTAGTGGAAGAAACTAGTTTTATTGATGCTGGTGCACCAACATCTGT
ACCTTCCATTCCCCCAGATGTATCAGGATTTAGTATTACTACTTCAACTGATACCACAAGCTTGCG

Block 2 (948 nt)

GGCAGATCTCTGGAGCTATATTAATACTATTATCATTACTAGAAAAATATAATGTATTATCCACATCTATACCTT
CATATGCAGGATTATCATATGTAATAAGTTTAGTGGGAGTGGTTATAAAAGCAGACTAGTGCCAGATCTTCCTGA
CTTTTTGGATATAGTTGCTTTACATAGGCCAGCATTAACCTCTAGGCGTACTGGCATTAGGTACAGTAGAATTGG
TAATAAACAAACACTACGTACTCGTAGTGGAAAATCTATACTAGTTGCAGATCTCCTTTGCCCCAGTGTTCCCCT
ATAGGTGGTTTGCAACCAATTAAACACAATTGTGTTTGTTTGTAATCCATAGATATACATTCTCTATTATCCACA
CCTGCATTTGCTGCATAAGCACTAACTAGTCCCAGATCTTCCCCATGTACCAATGTTGCAGTAAATCCAGGTGAT
TGTCCACCATTAGAGTTAATAAACACAGTTATTCAGGATGGTGATATGGTTGATACTGGCTTTGGTGCTATGGAC
TTTACTACAACTAGTCGGAGATCTCTTCTTTAGGTGCTGGAGGTGTATGTTTTTGACAAGCAATTGCCTGGGATG
TTACAAACCTATAAGTATCTTCTAGTGTGCCTCCTGGGGGAGGTTGTAGACCAAAATTCCAGTCCTCCAACTAGT
GGCAGATCTTCAGTTTCCTTTAGGACGCAAATTTTTACTACAAGCAGGATTGAAGGCCAAACCAAAATTTACATT
AGGAAAACGAAAAGCTACACCCACCACCTCATCTACCTCTACAACTGCTAAACGACTAGTGTCAGATCTCTGCAA
CAAGACATACATCGACCGGTCCACCGACCCCTTATATTATGGAATCTTTGCTTTTTGTCCAGATGTCTTTGCTTT
TCTTCAGGACACAGTGGCTTTTGACAGTTAATACACCTAACTAGTCGC

References

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., . . . Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature, 500*(7461), 207-211. doi: 10.1038/nature12064

Akagi, K., Li, J., Broutian, T. R., Padilla-Nash, H., Xiao, W., Jiang, B., . . . Gillison, M. L. (2014). Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res, 24*(2), 185-199. doi: 10.1101/gr.164806.113

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell, 16*(1), 47-57. doi: 10.1016/j.devcel.2008.11.011

Amaral, P. P., & Mattick, J. S. (2008). Noncoding RNA in development. *Mamm Genome, 19*(7-8), 454-492. doi: 10.1007/s00335-008-9136-7

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., . . . Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science, 340*(6137), 1234167. doi: 10.1126/science.1234167

Andrulis, E. D., Neiman, A. M., Zappulla, D. C., & Sternglanz, R. (1998). Perinuclear localization of chromatin facilitates transcriptional silencing. *Nature, 394*(6693), 592-595. doi: 10.1038/29100

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell, 27*(2 Pt 1), 299-308.

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res, 21*(3), 381-395. doi: 10.1038/cr.2011.22

Bannister, A. J., Schneider, R., Myers, F. A., Thorne, A. W., Crane-Robinson, C., & Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J Biol Chem, 280*(18), 17732-17736. doi: 10.1074/jbc.M500796200

Bannister, A. J., Zegerman, P., Partridge, J. F., Miska, E. A., Thomas, J. O., Allshire, R. C., & Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature, 410*(6824), 120-124. doi: 10.1038/35065138

Barozzi, P., Potenza, L., Riva, G., Vallerini, D., Quadrelli, C., Bosco, R., . . . Luppi, M. (2007). B cells and herpesviruses: a model of lymphoproliferation. *Autoimmun Rev, 7*(2), 132-136. doi: 10.1016/j.autrev.2007.02.018

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., . . . Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell, 129*(4), 823-837. doi: 10.1016/j.cell.2007.05.009

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods, 58*(3), 268-276. doi: 10.1016/j.ymeth.2012.05.001

Benedetti, F., Dorier, J., Burnier, Y., & Stasiak, A. (2014). Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic Acids Res, 42*(5), 2848-2855. doi: 10.1093/nar/gkt1353

Benedict, C. A., Angulo, A., Patterson, G., Ha, S., Huang, H., Messerle, M., . . . Ghazal, P. (2004). Neutrality of the canonical NF-kappaB-dependent pathway for human and murine cytomegalovirus transcription and replication in vitro. *J Virol, 78*(2), 741-750.

Bensaude, O. (2011). Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription, 2*(3), 103-108. doi: 10.4161/trns.2.3.16172

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. *Nature, 447*(7143), 407-412. doi: 10.1038/nature05915

Bernard, H. U. (2002). Gene expression of genital human papillomaviruses and considerations on potential antiviral approaches. *Antivir Ther, 7*(4), 219-237.

Bernard, H. U. (2013). Regulatory elements in the viral genome. *Virology, 445*(1-2), 197-204. doi: 10.1016/j.virol.2013.04.035

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*: Kluwer Academic Publishers.

Bi, X., Yu, Q., Sandmeier, J. J., & Zou, Y. (2004). Formation of boundaries of transcriptionally silent chromatin by nucleosome-excluding structures. *Mol Cell Biol, 24*(5), 2118-2131.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., . . . Canaider, S. (2013). An estimation of the number of cells in the human body. *Ann Hum Biol, 40*(6), 463-471. doi: 10.3109/03014460.2013.807878

Bickmore, W. A. (2013). The spatial organization of the human genome. *Annu Rev Genomics Hum Genet, 14*, 67-84. doi: 10.1146/annurev-genom-091212-153515

Bodelon, C., Untereiner, M. E., Machiela, M. J., Vinokurova, S., & Wentzensen, N. (2016). Genomic characterization of viral integration sites in HPV-related cancers. *Int J Cancer, 139*(9), 2001-2011. doi: 10.1002/ijc.30243

Bodily, J. M., Mehta, K. P., & Laimins, L. A. (2011). Human papillomavirus E7 enhances hypoxia-inducible factor 1-mediated transcription by inhibiting binding of histone deacetylases. *Cancer Res, 71*(3), 1187-1195. doi: 10.1158/0008-5472.can-10-2626

Boehmer, P. E., & Nimonkar, A. V. (2003). Herpes virus replication. *IUBMB Life, 55*(1), 13-22. doi: 10.1080/1521654031000070645

Boettiger, A. N., Bintu, B., Moffitt, J. R., Wang, S., Beliveau, B. J., Fudenberg, G., . . . Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature, 529*(7586), 418-422. doi: 10.1038/nature16496

Bolger, G., Lapeyre, N., Rheaume, M., Kibler, P., Bousquet, C., Garneau, M., & Cordingley, M. (1999). Acute murine cytomegalovirus infection: a model for determining antiviral activity against CMV induced hepatitis. *Antiviral Res, 44*(3), 155-165.

Bolland, D. J., King, M. R., Reik, W., Corcoran, A. E., & Krueger, C. (2013). Robust 3D DNA FISH using directly labeled probes. *J Vis Exp*(78). doi: 10.3791/50587

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., . . . Cremer, T. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol, 3*(5), e157. doi: 10.1371/journal.pbio.0030157

Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nat Rev Genet, 17*(12), 772. doi: 10.1038/nrg.2016.147

Bonnans, C., Chou, J., & Werb, Z. (2014). Remodelling the extracellular matrix in development and disease. *Nat Rev Mol Cell Biol, 15*(12), 786-801. doi: 10.1038/nrm3904

Borggrefe, T., & Yue, X. (2011). Interactions between subunits of the Mediator complex with gene-specific transcription factors. *Semin Cell Dev Biol, 22*(7), 759-768. doi: 10.1016/j.semcdb.2011.07.022

Bosse, J. B., Hogue, I. B., Feric, M., Thiberge, S. Y., Sodeik, B., Brangwynne, C. P., & Enquist, L. W. (2015). Remodeling nuclear architecture allows efficient transport of herpesvirus capsids by diffusion. *Proc Natl Acad Sci U S A, 112*(42), E5725-5733. doi: 10.1073/pnas.1513876112

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., . . . Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell, 132*(2), 311-322. doi: 10.1016/j.cell.2007.12.014

Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R., & Nasmyth, K. (1985). Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell, 41*(1), 41-48.

Brenowitz, M., Senear, D. F., & Kingston, R. E. (2001). DNase I footprint analysis of protein-DNA binding. *Curr Protoc Mol Biol, Chapter 12*, Unit 12.14. doi: 10.1002/0471142727.mb1204s07

Brent, R., & Ptashne, M. (1985). A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell, 43*(3 Pt 2), 729-736.

Bresnahan, W. A., & Shenk, T. (2000). A subset of viral transcripts packaged within human cytomegalovirus particles. *Science, 288*(5475), 2373-2376.

Brody, A. R., & Craighead, J. E. (1974). Pathogenesis of pulmonary cytomegalovirus infection in immunosuppressed mice. *J Infect Dis, 129*(6), 677-689.

Brown, S. W. (1966). Heterochromatin. *Science, 151*(3709), 417-425.

Browne, E. P., Wing, B., Coleman, D., & Shenk, T. (2001). Altered cellular mRNA levels in human cytomegalovirus-infected fibroblasts: viral block to the accumulation of antiviral mRNAs. *J Virol, 75*(24), 12319-12330. doi: 10.1128/jvi.75.24.12319-12330.2001

Brune, W., Nevels, M., & Shenk, T. (2003). Murine cytomegalovirus m41 open reading frame encodes a Golgi-localized antiapoptotic protein. *J Virol, 77*(21), 11633-11643.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods, 10*(12), 1213-1218. doi: 10.1038/nmeth.2688

Bulger, M., & Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell, 144*(3), 327-339. doi: 10.1016/j.cell.2011.01.024

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. *Mol Cell, 36*(4), 541-546. doi: 10.1016/j.molcel.2009.10.019

Burger, K., Muhl, B., Harasim, T., Rohrmoser, M., Malamoussi, A., Orban, M., . . . Eick, D. (2010). Chemotherapeutic drugs inhibit ribosome biogenesis at various levels. *J Biol Chem, 285*(16), 12416-12425. doi: 10.1074/jbc.M109.074211

Burk, R. (2017). Integrated genomic and molecular characterization of cervical cancer. *Nature, 543*(7645), 378-384. doi: 10.1038/nature21386

Caposio, P., Luganini, A., Bronzini, M., Landolfo, S., & Gribaudo, G. (2010). The Elk-1 and serum response factor binding sites in the major immediate-early promoter of human cytomegalovirus are required for efficient viral replication in quiescent cells and compensate for inactivation of the NF-kappaB sites in proliferating cells. *J Virol, 84*(9), 4481-4493. doi: 10.1128/jvi.02141-09

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., . . . Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet, 38*(6), 626-635. doi: 10.1038/ng1789

Carrozza, M. J., Li, B., Florens, L., Suganuma, T., Swanson, S. K., Lee, K. K., . . . Workman, J. L. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell, 123*(4), 581-592. doi: 10.1016/j.cell.2005.10.023

Castillo, J. P., Frame, F. M., Rogoff, H. A., Pickering, M. T., Yurochko, A. D., & Kowalik, T. F. (2005). Human cytomegalovirus IE1-72 activates ataxia telangiectasia mutated kinase and a p53/p21-mediated growth arrest response. *J Virol, 79*(17), 11467-11475. doi: 10.1128/jvi.79.17.11467-11475.2005

Challacombe, J. F., Rechtsteiner, A., Gottardo, R., Rocha, L. M., Browne, E. P., Shenk, T., . . . Brettin, T. S. (2004). Evaluation of the host transcriptional response to human cytomegalovirus infection. *Physiol Genomics, 18*(1), 51-62. doi: 10.1152/physiolgenomics.00155.2003

Chambeyron, S., & Bickmore, W. A. (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev, 18*(10), 1119-1130. doi: 10.1101/gad.292104

Chan, B., & Goncalves Magalhaes, V. (2017). The murine cytomegalovirus M35 protein antagonizes type I IFN induction downstream of pattern recognition receptors by targeting NF-kappaB mediated transcription. *13*(5), e1006382. doi: 10.1371/journal.ppat.1006382

Chandra, T., Ewels, P. A., Schoenfelder, S., Furlan-Magaril, M., Wingett, S. W., Kirschner, K., . . . Reik, W. (2015). Global reorganization of the nuclear landscape in senescent cells. *Cell Rep, 10*(4), 471-483. doi: 10.1016/j.celrep.2014.12.055

Chang, H. R., Munkhjargal, A., Kim, M. J., Park, S. Y., Jung, E., Ryu, J. H., . . . Kim, Y. (2017). The functional roles of PML nuclear bodies in genome maintenance. *Mutat Res*. doi: 10.1016/j.mrfmmm.2017.05.002

Chaumeil, J., Le Baccon, P., Wutz, A., & Heard, E. (2006). A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev, 20*(16), 2223-2237. doi: 10.1101/gad.380906

Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., . . . et al. (1990). Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Curr Top Microbiol Immunol, 154*, 125-169.

Cheeran, M. C., Lokensgard, J. R., & Schleiss, M. R. (2009). Neuropathogenesis of congenital cytomegalovirus infection: disease mechanisms and prospects for intervention. *Clin Microbiol Rev, 22*(1), 99-126, Table of Contents. doi: 10.1128/cmr.00023-08

Chen, Y., Williams, V., Filippova, M., Filippov, V., & Duerksen-Hughes, P. (2014). Viral carcinogenesis: factors inducing DNA damage and virus integration. *Cancers (Basel), 6*(4), 2155-2186. doi: 10.3390/cancers6042155

Chesterton, C. J., Coupar, B. E., & Butterworth, P. H. (1974). Transcription of fractionated mammalian chromatin by mammalian ribonucleic acid polymerase. Demonstration of temperature-dependent rifampicin-resistant initiation sites in euchromatin deoxyribonucleic acid. *Biochem J, 143*(1), 73-81.

Christiansen, I. K., Sandve, G. K., Schmitz, M., Durst, M., & Hovig, E. (2015). Transcriptionally active regions are the preferred targets for chromosomal HPV integration in cervical carcinogenesis. *PLoS One, 10*(3), e0119566. doi: 10.1371/journal.pone.0119566

Churchman, L. S., & Weissman, J. S. (2012). Native elongating transcript sequencing (NET-seq). *Curr Protoc Mol Biol, Chapter 4*, Unit 4.14.11-17. doi: 10.1002/0471142727.mb0414s98

Comet, I., Schuettengruber, B., Sexton, T., & Cavalli, G. (2011). A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. *Proc Natl Acad Sci U S A, 108*(6), 2294-2299. doi: 10.1073/pnas.1002059108

Compton, T., & Feire, A. (2007). Early events in human cytomegalovirus infection. In A. Arvin, G. Campadelli-Fiume, E. Mocarski, P. S. Moore, B. Roizman, R. Whitley & K. Yamanishi (Eds.), *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*. Cambridge: Cambridge University Press.

Conger, K. L., Liu, J. S., Kuo, S. R., Chow, L. T., & Wang, T. S. (1999). Human papillomavirus DNA replication. Interactions between the viral E1 protein and two subunits of human dna polymerase alpha/primase. *J Biol Chem, 274*(5), 2696-2705.

Cook, P. R. (1999). The organization of replication and transcription. *Science, 284*(5421), 1790-1795.

Cook, P. R. (2002). Predicting three-dimensional genome structure from transcriptional activity. *Nat Genet, 32*(3), 347-352. doi: 10.1038/ng1102-347

Costantini, M., Cammarano, R., & Bernardi, G. (2009). The evolution of isochore patterns in vertebrate genomes. *BMC Genomics, 10*, 146. doi: 10.1186/1471-2164-10-146

Cremer, T., & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet, 2*(4), 292-301. doi: 10.1038/35066075

Cremer, T., Cremer, C., Baumann, H., Luedtke, E. K., Sperling, K., Teuber, V., & Zorn, C. (1982). Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments. *Hum Genet, 60*(1), 46-56.

Crosbie, E. J., Einstein, M. H., Franceschi, S., & Kitchener, H. C. (2013). Human papillomavirus and cervical cancer. *Lancet, 382*(9895), 889-899. doi: 10.1016/s0140-6736(13)60022-7

Csink, A. K., & Henikoff, S. (1996). Genetic modification of heterochromatic association and nuclear organization in Drosophila. *Nature, 381*(6582), 529-531. doi: 10.1038/381529a0

Dall, K. L., Scarpini, C. G., Roberts, I., Winder, D. M., Stanley, M. A., Muralidhar, B., . . . Coleman, N. (2008). Characterization of naturally occurring HPV16 integration sites isolated from cervical keratinocytes under noncompetitive conditions. *Cancer Res, 68*(20), 8249-8259. doi: 10.1158/0008-5472.can-08-1741

Damato, E. G., & Winnen, C. W. (2002). Cytomegalovirus infection: perinatal implications. *J Obstet Gynecol Neonatal Nurs, 31*(1), 86-92.

Datta, S., & Datta, S. (2006). Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics, 7 Suppl 4*, S17. doi: 10.1186/1471-2105-7-s4-s17

Datta, S. D., & Saraiya, M. (2011). Cervical cancer screening among women who attend sexually transmitted diseases (STD) clinics: background paper for 2010 STD Treatment Guidelines. *Clin Infect Dis, 53 Suppl 3*, S153-159. doi: 10.1093/cid/cir704

Davies, J. O., Oudelaar, A. M., Higgs, D. R., & Hughes, J. R. (2017). How best to identify chromosomal interactions: a comparison of approaches. *14*(2), 125-134. doi: 10.1038/nmeth.4146

Davison, A. J., Eberle, R., Ehlers, B., Hayward, G. S., McGeoch, D. J., Minson, A. C., . . . Thiry, E. (2009). The order Herpesvirales. *Arch Virol, 154*(1), 171-177. doi: 10.1007/s00705-008-0278-4

de Laat, W., & Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature, 502*(7472), 499-506. doi: 10.1038/nature12753

de Villiers, E. M., Fauquet, C., Broker, T. R., Bernard, H. U., & zur Hausen, H. (2004). Classification of papillomaviruses. *Virology, 324*(1), 17-27. doi: 10.1016/j.virol.2004.03.033

de Wit, E., Bouwman, B. A., Zhu, Y., Klous, P., Splinter, E., Verstegen, M. J., . . . de Laat, W. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature, 501*(7466), 227-231. doi: 10.1038/nature12420

de Wit, E., & de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes Dev, 26*(1), 11-24. doi: 10.1101/gad.179804.111

de Wit, E., Greil, F., & van Steensel, B. (2007). High-resolution mapping reveals links of HP1 with active and inactive chromatin components. *PLoS Genet, 3*(3), e38. doi: 10.1371/journal.pgen.0030038

de Wit, E., Vos, E. S., Holwerda, S. J., Valdes-Quezada, C., Verstegen, M. J., Teunissen, H., . . . de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Mol Cell, 60*(4), 676-684. doi: 10.1016/j.molcel.2015.09.023

Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet, 14*(6), 390-403. doi: 10.1038/nrg3454

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science, 295*(5558), 1306-1311. doi: 10.1126/science.1067799

Deng, W., Rupon, J. W., Krivega, I., Breda, L., Motta, I., Jahn, K. S., . . . Blobel, G. A. (2014). Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell, 158*(4), 849-860. doi: 10.1016/j.cell.2014.05.050

Dernburg, A. F., Broman, K. W., Fung, J. C., Marshall, W. F., Philips, J., Agard, D. A., & Sedat, J. W. (1996). Perturbation of nuclear architecture by long-distance chromosome interactions. *Cell, 85*(5), 745-759.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., & Pagano, A. (2007). The expanding RNA polymerase III transcriptome. *Trends Genet, 23*(12), 614-622. doi: 10.1016/j.tig.2007.09.001

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., . . . Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature, 518*(7539), 331-336. doi: 10.1038/nature14222

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., . . . Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature, 485*(7398), 376-380. doi: 10.1038/nature11082

Dolken, L., Perot, J., Cognat, V., Alioua, A., John, M., Soutschek, J., . . . Pfeffer, S. (2007). Mouse cytomegalovirus microRNAs dominate the cellular small RNA profile during lytic infection and show features of posttranscriptional regulation. *J Virol, 81*(24), 13771-13782. doi: 10.1128/jvi.01313-07

Doolittle-Hall, J. M., Cunningham Glasspoole, D. L., Seaman, W. T., & Webster-Cyriaque, J. (2015). Meta-Analysis of DNA Tumor-Viral Integration Site Selection Indicates a Role for Repeats, Gene Expression and Epigenetics. *Cancers (Basel), 7*(4), 2217-2235. doi: 10.3390/cancers7040887

Doorbar, J. (2005). The papillomavirus life cycle. *J Clin Virol, 32 Suppl 1*, S7-15. doi: 10.1016/j.jcv.2004.12.006

Doorbar, J., Quint, W., Banks, L., Bravo, I. G., Stoler, M., Broker, T. R., & Stanley, M. A. (2012). The biology and life-cycle of human papillomaviruses. *Vaccine, 30 Suppl 5*, F55-70. doi: 10.1016/j.vaccine.2012.06.083

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., . . . Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res, 16*(10), 1299-1309. doi: 10.1101/gr.5571506

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., . . . Young, R. A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell, 159*(2), 374-387. doi: 10.1016/j.cell.2014.09.030

Draghici, S., Khatri, P., Eklund, A. C., & Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet, 22*(2), 101-109. doi: 10.1016/j.tig.2005.12.005

Dryden, N. H., Broome, L. R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., . . . Fletcher, O. (2014). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res, 24*(11), 1854-1868. doi: 10.1101/gr.175034.114

Duensing, S., & Munger, K. (2002). Human papillomaviruses and centrosome duplication errors: modeling the origins of genomic instability. *Oncogene, 21*(40), 6241-6248. doi: 10.1038/sj.onc.1205709

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3*(3), 32-57. doi: 10.1080/01969727308546046

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016a). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst, 3*(1), 99-101. doi: 10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016b). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst, 3*(1), 95-98. doi: 10.1016/j.cels.2016.07.002

Durst, M., Gissmann, L., Ikenberg, H., & zur Hausen, H. (1983). A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. *Proc Natl Acad Sci U S A, 80*(12), 3812-3815.

Engreitz, J., Lander, E. S., & Guttman, M. (2015). RNA antisense purification (RAP) for mapping RNA interactions with chromatin. *Methods Mol Biol, 1262*, 183-197. doi: 10.1007/978-1-4939-2253-6_11

Engstrom, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S., & Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res, 17*(12), 1898-1908. doi: 10.1101/gr.6669607

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., . . . Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer, 136*(5), E359-386. doi: 10.1002/ijc.29210

Fiers, W., Contreras, R., Haegemann, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., . . . Ysebaert, M. (1978). Complete nucleotide sequence of SV40 DNA. *Nature, 273*(5658), 113-120.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., . . . van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell, 143*(2), 212-224. doi: 10.1016/j.cell.2010.09.009

Fliss, P. M., Jowers, T. P., Brinkmann, M. M., Holstermann, B., Mack, C., Dickinson, P., . . . Brune, W. (2012). Viral mediated redirection of NEMO/IKKgamma to autophagosomes curtails the inflammatory cascade. *PLoS Pathog, 8*(2), e1002517. doi: 10.1371/journal.ppat.1002517

Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., & Ferrari, F. (2017). Comparison of computational methods for Hi-C data analysis. doi: 10.1038/nmeth.4325

Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., Haberle, V., . . . Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature, 507*(7493), 462-470. doi: 10.1038/nature13182

Fortin, J. P., & Hansen, K. D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol, 16*, 180. doi: 10.1186/s13059-015-0741-y

Freire-Pritchett, P., Schoenfelder, S., Varnai, C., Wingett, S. W., Cairns, J., Collier, A. J., . . . Spivakov, M. (2017). Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *6*. doi: 10.7554/eLife.21926

Fudenberg, G., & Imakaev, M. (2017). FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat Methods, 14*(7), 673-678. doi: 10.1038/nmeth.4329

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep, 15*(9), 2038-2049. doi: 10.1016/j.celrep.2016.04.085

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., . . . Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature, 462*(7269), 58-64. doi: 10.1038/nature08497

Fulop, T., Larbi, A., & Pawelec, G. (2013). Human T cell aging and the impact of persistent viral infections. *Front Immunol, 4*, 271. doi: 10.3389/fimmu.2013.00271

Futschik, M. E., & Carlisle, B. (2005). Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol, 3*(4), 965-988.

Gagen, M. J., & Mattick, J. S. (2005). Inherent size constraints on prokaryote gene networks due to "accelerating" growth. *Theory Biosci, 123*(4), 381-411. doi: 10.1016/j.thbio.2005.02.002

Gall, J. G. (2003). The centennial of the Cajal body. *Nat Rev Mol Cell Biol, 4*(12), 975-980. doi: 10.1038/nrm1262

Gandhi, M. K., & Khanna, R. (2004). Human cytomegalovirus: clinical aspects, immune regulation, and emerging treatments. *Lancet Infect Dis, 4*(12), 725-738. doi: 10.1016/s1473-3099(04)01202-2

Gaspar, M., & Shenk, T. (2006). Human cytomegalovirus inhibits a DNA damage response by mislocalizing checkpoint proteins. *Proc Natl Acad Sci U S A, 103*(8), 2821-2826. doi: 10.1073/pnas.0511148103

Gaszner, M., & Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet, 7*(9), 703-713. doi: 10.1038/nrg1925

Gat-Viks, I., Sharan, R., & Shamir, R. (2003). Scoring clustering solutions by their biological relevance. *Bioinformatics, 19*(18), 2381-2389.

Gay, L., Karfilis, K. V., Miller, M. R., Doe, C. Q., & Stankunas, K. (2014). Applying thiouracil tagging to mouse transcriptome analysis. *Nat Protoc, 9*(2), 410-420. doi: 10.1038/nprot.2014.023

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K. K., Cheng, C., . . . Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature, 489*(7414), 91-100. doi: 10.1038/nature11245

Gibbs, H. C., Sing, G., Gonzalez Armas, J. C., Campbell, C. J., Ghazal, P., & Yeh, A. T. (2013). Time-lapse ultrashort pulse microscopy of infection in three-dimensional versus two-dimensional culture environments reveals enhanced extra-chromosomal virus replication compartment formation. *J Biomed Opt, 18*(3), 031111. doi: 10.1117/1.jbo.18.3.031111

Gibcus, J. H., & Dekker, J. (2013). The hierarchy of the 3D genome. *Mol Cell, 49*(5), 773-782. doi: 10.1016/j.molcel.2013.02.011

Gibeault, R. L., & Conn, K. L. (2016). An Essential Viral Transcription Activator Modulates Chromatin Dynamics. *12*(8), e1005842. doi: 10.1371/journal.ppat.1005842

Gibson, W. (1996). Structure and assembly of the virion. *Intervirology, 39*(5-6), 389-400.

Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., Dekker, J., . . . Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell, 157*(4), 950-963. doi: 10.1016/j.cell.2014.03.025

Giorgetti, L., & Heard, E. (2016). Closing the loop: 3C versus DNA FISH. *Genome Biol, 17*(1), 215. doi: 10.1186/s13059-016-1081-2

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., . . . Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol, 27*(2), 182-189. doi: 10.1038/nbt.1523

Goldmacher, V. S., Bartle, L. M., Skaletskaya, A., Dionne, C. A., Kedersha, N. L., Vater, C. A., . . . Chittenden, T. (1999). A cytomegalovirus-encoded mitochondria-localized inhibitor of apoptosis structurally unrelated to Bcl-2. *Proc Natl Acad Sci U S A, 96*(22), 12536-12541.

Gorkin, D. U., Leung, D., & Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell, 14*(6), 762-775. doi: 10.1016/j.stem.2014.05.017

Grande, M. A., van der Kraan, I., de Jong, L., & van Driel, R. (1997). Nuclear distribution of transcription factors in relation to sites of transcription and RNA polymerase II. *J Cell Sci, 110 (Pt 15)*, 1781-1791.

Grau, D. J., Chapman, B. A., Garlick, J. D., Borowsky, M., Francis, N. J., & Kingston, R. E. (2011). Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. *Genes Dev, 25*(20), 2210-2221. doi: 10.1101/gad.17288211

Groves, I. J., & Coleman, N. (2015). Pathogenesis of human papillomavirus-associated mucosal disease. *J Pathol, 235*(4), 527-538. doi: 10.1002/path.4496

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., . . . van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature, 453*(7197), 948-951. doi: 10.1038/nature06947

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., . . . Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell, 141*(1), 129-141. doi: 10.1016/j.cell.2010.03.009

Hahn, M. W., & Wray, G. A. (2002). The g-value paradox. *Evol Dev, 4*(2), 73-75.

Hahn, W. C. (2002). Immortalization and transformation of human cells. *Mol Cells, 13*(3), 351-361.

Hall, L. L., Carone, D. M., Gomez, A. V., Kolpa, H. J., Byron, M., Mehta, N., . . . Lawrence, J. B. (2014). Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell, 156*(5), 907-919. doi: 10.1016/j.cell.2014.01.042

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell, 100*(1), 57-70.

Harewood, L., Kishore, K., Eldridge, M. D., Wingett, S., Pearson, D., Schoenfelder, S., . . . Fraser, P. (2017). Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *18*(1), 125. doi: 10.1186/s13059-017-1253-8

Harmon, B., & Sedat, J. (2005). Cell-by-cell dissection of gene expression and chromosomal interactions reveals consequences of nuclear reorganization. *PLoS Biol, 3*(3), e67. doi: 10.1371/journal.pbio.0030067

Hatano, T., Sano, D., Takahashi, H., Hyakusoku, H., Isono, Y., Shimada, S., . . . Oridate, N. (2017). Identification of human papillomavirus (HPV) 16 DNA integration and the ensuing patterns of methylation in HPV-associated head and neck squamous cell carcinoma cell lines. *Int J Cancer, 140*(7), 1571-1580. doi: 10.1002/ijc.30589

Heiman, M., Kulicke, R., Fenster, R. J., Greengard, P., & Heintz, N. (2014). Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). *Nat Protoc, 9*(6), 1282-1291. doi: 10.1038/nprot.2014.085

Heitz, E. (1928). Das Heterochromatin der Moose. *I Jahrb Wiss Botanik, 69*, 762 - 818. doi: citeulike-article-id:5963590

Heming, J. D., Conway, J. F., & Homa, F. L. (2017). Herpesvirus Capsid Assembly and DNA Packaging. *Adv Anat Embryol Cell Biol, 223*, 119-142. doi: 10.1007/978-3-319-53168-7_6

Hendzel, M. J., Kruhlak, M. J., & Bazett-Jones, D. P. (1998). Organization of highly acetylated chromatin around sites of heterogeneous nuclear RNA accumulation. *Mol Biol Cell, 9*(9), 2491-2507.

Hertel, L., Chou, S., & Mocarski, E. S. (2007). Viral and cell cycle-regulated kinases in cytomegalovirus-induced pseudomitosis and replication. *PLoS Pathog, 3*(1), e6. doi: 10.1371/journal.ppat.0030006

Hertel, L., & Mocarski, E. S. (2004). Global analysis of host cell gene expression late during cytomegalovirus infection reveals extensive dysregulation of cell cycle gene expression and induction of Pseudomitosis independent of US28 function. *J Virol, 78*(21), 11988-12011. doi: 10.1128/jvi.78.21.11988-12011.2004

Hickman, E. S., Moroni, M. C., & Helin, K. (2002). The role of p53 and pRB in apoptosis and cancer. *Curr Opin Genet Dev, 12*(1), 60-66.

Hiratani, I., Ryba, T., Itoh, M., Rathjen, J., Kulik, M., Papp, B., . . . Gilbert, D. M. (2010). Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res, 20*(2), 155-169. doi: 10.1101/gr.099796.109

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A. L., Bak, R. O., Li, C. H., . . . Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science, 351*(6280), 1454-1458. doi: 10.1126/science.aad9024

Holmes, A., Lameiras, S., Jeannot, E., Marie, Y., Castera, L., Sastre-Garau, X., & Nicolas, A. (2016). Mechanistic signatures of HPV insertions in cervical carcinomas. *1*, 16004. doi: 10.1038/npjgenmed.2016.4

Hu, Z., Zhu, D., Wang, W., Li, W., Jia, W., Zeng, X., . . . Ma, D. (2015). Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet, 47*(2), 158-163. doi: 10.1038/ng.3178

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc, 4*(1), 44-57. doi: 10.1038/nprot.2008.211

Huang, J., Kent, J. R., Placek, B., Whelan, K. A., Hollow, C. M., Zeng, P. Y., . . . Berger, S. L. (2006). Trimethylation of histone H3 lysine 4 by Set1 in the lytic infection of human herpes simplex virus 1. *J Virol, 80*(12), 5740-5746. doi: 10.1128/jvi.00169-06

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., . . . Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet, 46*(2), 205-212. doi: 10.1038/ng.2871

Hwang, J., & Kalejta, R. F. (2007). Proteasome-dependent, ubiquitin-independent degradation of Daxx by the viral pp71 protein in human cytomegalovirus-infected cells. *Virology, 367*(2), 334-338. doi: 10.1016/j.virol.2007.05.037

Iborra, F. J., Pombo, A., Jackson, D. A., & Cook, P. R. (1996). Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. *J Cell Sci, 109 (Pt 6)*, 1427-1436.

Ihalainen, T. O., Niskanen, E. A., Jylhava, J., Paloheimo, O., Dross, N., Smolander, H., . . . Vihinen-Ranta, M. (2009). Parvovirus induced alterations in nuclear architecture and dynamics. *PLoS One, 4*(6), e5948. doi: 10.1371/journal.pone.0005948

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., . . . Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods, 9*(10), 999-1003. doi: 10.1038/nmeth.2148

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science, 324*(5924), 218-223. doi: 10.1126/science.1168978

Ioudinkova, E., Arcangeletti, M. C., Rynditch, A., De Conto, F., Motta, F., Covan, S., . . . Chezzi, C. (2006). Control of human cytomegalovirus gene expression by differential histone modifications during lytic and latent infection of a monocytic cell line. *Gene, 384*, 120-128. doi: 10.1016/j.gene.2006.07.021

Isaacson, M. K., Juckem, L. K., & Compton, T. (2008). Virus entry and innate immune activation. *Curr Top Microbiol Immunol, 325*, 85-100.

Jackson, D. A., Hassan, A. B., Errington, R. J., & Cook, P. R. (1993). Visualization of focal sites of transcription within human nuclei. *Embo j, 12*(3), 1059-1065.

Jager, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H. E., Heindl, A., . . . Houlston, R. S. (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun, 6*, 6178. doi: 10.1038/ncomms7178

Jault, F. M., Jault, J. M., Ruchti, F., Fortunato, E. A., Clark, C., Corbeil, J., . . . Spector, D. H. (1995). Cytomegalovirus infection induces high levels of cyclins, phosphorylated Rb, and p53, leading to cell cycle arrest. *J Virol, 69*(11), 6697-6704.

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., . . . Fraser, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell, 167*(5), 1369-1384.e1319. doi: 10.1016/j.cell.2016.09.037

Jiang, C., & Pugh, B. F. (2009). Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet, 10*(3), 161-172. doi: 10.1038/nrg2522

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., . . . Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature, 503*(7475), 290-294. doi: 10.1038/nature12644

Jonkers, I., & Lis, J. T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol, 16*(3), 167-177. doi: 10.1038/nrm3953

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., . . . Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell, 128*(6), 1231-1245. doi: 10.1016/j.cell.2006.12.048

Kimura, A., & Horikoshi, M. (2004). Partition of distinct chromosomal regions: negotiable border and fixed border. *Genes Cells, 9*(6), 499-508. doi: 10.1111/j.1356-9597.2004.00740.x

Kimura, A., Umehara, T., & Horikoshi, M. (2002). Chromosomal gradient of histone acetylation established by Sas2p and Sir2p functions as a shield against gene silencing. *Nat Genet, 32*(3), 370-377. doi: 10.1038/ng993

Kind, J., & van Steensel, B. (2010). Genome-nuclear lamina interactions and gene regulation. *Curr Opin Cell Biol, 22*(3), 320-325. doi: 10.1016/j.ceb.2010.04.002

Klein, F. A., Pakozdi, T., Anders, S., Ghavi-Helm, Y., Furlong, E. E., & Huber, W. (2015). FourCSeq: analysis of 4C sequencing data. *Bioinformatics, 31*(19), 3085-3091. doi: 10.1093/bioinformatics/btv335

Kleinjan, D. A., & Lettice, L. A. (2008). Long-range gene control and genetic disease. *Adv Genet, 61*, 339-388. doi: 10.1016/s0065-2660(07)00013-2

Klemola, E., Von Essen, R., Henle, G., & Henle, W. (1970). Infectious-mononucleosis-like disease with negative heterophil agglutination test. Clinical features in relation to Epstein-Barr virus and cytomegalovirus antibodies. *J Infect Dis, 121*(6), 608-614.

Kobiler, O., Brodersen, P., Taylor, M. P., Ludmir, E. B., & Enquist, L. W. (2011). Herpesvirus replication compartments originate with single incoming viral genomes. *MBio, 2*(6). doi: 10.1128/mBio.00278-11

Kobiler, O., Lipman, Y., Therkelsen, K., Daubechies, I., & Enquist, L. W. (2010). Herpesviruses carrying a Brainbow cassette reveal replication and expression of limited numbers of incoming genomes. *Nat Commun, 1*, 146. doi: 10.1038/ncomms1145

Kolovos, P., Georgomanolis, T., Koeferle, A., Larkin, J. D., Brant, L., Nikolicc, M., . . . Papantonis, A. (2016). Binding of nuclear factor kappaB to noncanonical consensus sites reveals its multimodal role during the early inflammatory response. *Genome Res, 26*(11), 1478-1489. doi: 10.1101/gr.210005.116

Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., & Papantonis, A. (2012). Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin, 5*(1), 1. doi: 10.1186/1756-8935-5-1

Korbel, J. O., & Lee, C. (2013). Genome assembly and haplotyping with Hi-C. *Nat Biotechnol, 31*(12), 1099-1101. doi: 10.1038/nbt.2764

Kornberg, R. D. (1999). Eukaryotic transcriptional control. *Trends Cell Biol, 9*(12), M46-49.

Kosak, S. T., Skok, J. A., Medina, K. L., Riblet, R., Le Beau, M. M., Fisher, A. G., & Singh, H. (2002). Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science, 296*(5565), 158-162. doi: 10.1126/science.1068768

Krijger, P. H., & de Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. *Nat Rev Mol Cell Biol, 17*(12), 771-782. doi: 10.1038/nrm.2016.138

Krijger, P. H., & de Laat, W. (2017). Can We Just Say: Transcription Second? *Cell, 169*(2), 184-185. doi: 10.1016/j.cell.2017.03.026

Kropp, K. A., Simon, C. O., Fink, A., Renzaho, A., Kuhnapfel, B., Podlech, J., . . . Grzimek, N. K. (2009). Synergism between the components of the bipartite major immediate-early transcriptional enhancer of murine cytomegalovirus does not accelerate virus replication in cell culture and host tissues. *J Gen Virol, 90*(Pt 10), 2395-2401. doi: 10.1099/vir.0.012245-0

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res, 19*(9), 1639-1645. doi: 10.1101/gr.092759.109

Kumar, L., & M, E. F. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformation, 2*(1), 5-7.

Kuznetsova, T., Wang, S. Y., Rao, N. A., Mandoli, A., Martens, J. H., Rother, N., . . . Stunnenberg, H. G. (2015). Glucocorticoid receptor and nuclear factor kappa-b affect three-dimensional chromatin organization. *Genome Biol, 16*, 264. doi: 10.1186/s13059-015-0832-9

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., & Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature, 410*(6824), 116-120. doi: 10.1038/35065132

Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods, 72*, 65-75. doi: 10.1016/j.ymeth.2014.10.031

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi: 10.1038/35057062

Lane, D. P. (1992). Cancer. p53, guardian of the genome. *Nature, 358*(6381), 15-16. doi: 10.1038/358015a0

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi: 10.1038/nmeth.1923

Lauberth, S. M., Nakayama, T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H., & Roeder, R. G. (2013). H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell, 152*(5), 1021-1036. doi: 10.1016/j.cell.2013.01.052

Le Dily, F., Bau, D., Pohl, A., Vicent, G. P., Serra, F., Soronellas, D., . . . Beato, M. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev, 28*(19), 2151-2162. doi: 10.1101/gad.241422.114

Le, T. B., Imakaev, M. V., Mirny, L. A., & Laub, M. T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. *Science, 342*(6159), 731-734. doi: 10.1126/science.1242059

Lee, T. I., & Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet, 34*, 77-137. doi: 10.1146/annurev.genet.34.1.77

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell, 152*(6), 1237-1251. doi: 10.1016/j.cell.2013.02.014

Leibiger, C., Kosyakova, N., Mkrtchyan, H., Glei, M., Trifonov, V., & Liehr, T. (2013). First molecular cytogenetic high resolution characterization of the NIH 3T3 cell line by murine multicolor banding. *J Histochem Cytochem, 61*(4), 306-312. doi: 10.1369/0022155413476868

Lelli, K. M., Slattery, M., & Mann, R. S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. *Annu Rev Genet, 46*, 43-68. doi: 10.1146/annurev-genet-110711-155437

Lenhard, B., Sandelin, A., & Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet, 13*(4), 233-245. doi: 10.1038/nrg3163

Lettice, L. A., Williamson, I., Wiltshire, J. H., Peluso, S., Devenney, P. S., Hill, A. E., . . . Hill, R. E. (2012). Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. *Dev Cell, 22*(2), 459-467. doi: 10.1016/j.devcel.2011.12.010

Levine, A. J. (2009). The common mechanisms of transformation by the small DNA tumor viruses: The inactivation of tumor suppressor gene products: p53. *Virology, 384*(2), 285-293. doi: 10.1016/j.virol.2008.09.034

Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature, 424*(6945), 147-151. doi: 10.1038/nature01763

Li, Y., Danzer, J. R., Alvarez, P., Belmont, A. S., & Wallrath, L. L. (2003). Effects of tethering HP1 to euchromatic regions of the Drosophila genome. *Development, 130*(9), 1817-1824.

Liang, G., Lin, J. C., Wei, V., Yoo, C., Cheng, J. C., Nguyen, C. T., . . . Jones, P. A. (2004). Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc Natl Acad Sci U S A, 101*(19), 7357-7362. doi: 10.1073/pnas.0401866101

Liang, K., & Keles, S. (2012). Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics, 28*(1), 121-122. doi: 10.1093/bioinformatics/btr605

Liao, J. B. (2006). Viruses and human cancer. *Yale J Biol Med, 79*(3-4), 115-122.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326*(5950), 289-293. doi: 10.1126/science.1181369

Liesegang, T. J. (1992). Biology and molecular aspects of herpes simplex and varicella-zoster virus infections. *Ophthalmology, 99*(5), 781-799.

Liu, B., & Stinski, M. F. (1992). Human cytomegalovirus contains a tegument protein that enhances transcription from promoters with upstream ATF and AP-1 cis-acting elements. *J Virol, 66*(7), 4434-4444.

Liu, X., Bushnell, D. A., & Kornberg, R. D. (2013). RNA polymerase II transcription: structure and mechanism. *Biochim Biophys Acta, 1829*(1), 2-8. doi: 10.1016/j.bbagrm.2012.09.003

Liu, Y., Lu, Z., Xu, R., & Ke, Y. (2016). Comprehensive mapping of the human papillomavirus (HPV) DNA integration sites in cervical carcinomas by HPV capture technology. *Oncotarget, 7*(5), 5852-5864. doi: 10.18632/oncotarget.6809

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature, 389*(6648), 251-260. doi: 10.1038/38444

Lukac, D. M., Manuppello, J. R., & Alwine, J. C. (1994). Transcriptional activation by the human cytomegalovirus immediate-early proteins: requirements for simple promoter structures and interactions with multiple components of the transcription complex. *J Virol, 68*(8), 5184-5193.

Luo, L., Gassman, K. L., Petell, L. M., Wilson, C. L., Bewersdorf, J., & Shopland, L. S. (2009). The nuclear periphery of embryonic stem cells is a transcriptionally permissive and repressive compartment. *J Cell Sci, 122*(Pt 20), 3729-3737. doi: 10.1242/jcs.052555

Marcinowski, L., Lidschreiber, M., Windhager, L., Rieder, M., Bosse, J. B., Radle, B., . . . Dolken, L. (2012). Real-time transcriptional profiling of cellular and viral gene expression during lytic cytomegalovirus infection. *PLoS Pathog, 8*(9), e1002908. doi: 10.1371/journal.ppat.1002908

Martinez, F. P., Cosme, R. S., & Tang, Q. (2010). Murine cytomegalovirus major immediate-early protein 3 interacts with cellular and viral proteins in viral DNA replication compartments and is important for early gene activation. *J Gen Virol, 91*(Pt 11), 2664-2676. doi: 10.1099/vir.0.022301-0

Matharu, N., & Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer-Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. *PLoS Genet, 11*(12), e1005640. doi: 10.1371/journal.pgen.1005640

Mattick, J. S., & Gagen, M. J. (2005). Mathematics/computation. Accelerating networks. *Science, 307*(5711), 856-858. doi: 10.1126/science.1103737

Maul, G. G., & Negorev, D. (2008). Differences between mouse and human cytomegalovirus interactions with their respective hosts at immediate early times of the replication cycle. *Med Microbiol Immunol, 197*(2), 241-249. doi: 10.1007/s00430-008-0078-1

Mavrich, T. N., Jiang, C., Ioshikhes, I. P., Li, X., Venters, B. J., Zanton, S. J., . . . Pugh, B. F. (2008). Nucleosome organization in the Drosophila genome. *Nature, 453*(7193), 358-362. doi: 10.1038/nature06929

McBride, A. A., & Warburton, A. (2017). The role of integration in oncogenic progression of HPV-associated cancers. *PLoS Pathog, 13*(4), e1006211. doi: 10.1371/journal.ppat.1006211

McLaughlin-Drubin, M. E., & Munger, K. (2009). The human papillomavirus E7 oncoprotein. *Virology, 384*(2), 335-344. doi: 10.1016/j.virol.2008.10.006

Melo, C. A., Drost, J., Wijchers, P. J., van de Werken, H., de Wit, E., Oude Vrielink, J. A., . . . Agami, R. (2013). eRNAs are required for p53-dependent enhancer activity and gene transcription. *Mol Cell, 49*(3), 524-535. doi: 10.1016/j.molcel.2012.11.021

Mettenleiter, T. C. (2006). Intriguing interplay between viral proteins during herpesvirus assembly or: the herpesvirus assembly puzzle. *Vet Microbiol, 113*(3-4), 163-169. doi: 10.1016/j.vetmic.2005.11.040

Mettenleiter, T. C., Klupp, B. G., & Granzow, H. (2009). Herpesvirus assembly: an update. *Virus Res, 143*(2), 222-234. doi: 10.1016/j.virusres.2009.03.018

Michel, M., & Demel, C. (2017). TT-seq captures enhancer landscapes immediately after T-cell stimulation. *13*(3), 920. doi: 10.15252/msb.20167507

Miele, A., Bystricky, K., & Dekker, J. (2009). Yeast silent mating type loci form heterochromatic clusters through silencer protein-dependent long-range interactions. *PLoS Genet, 5*(5), e1000478. doi: 10.1371/journal.pgen.1000478

Mifsud, B., Martincorena, I., Darbo, E., Sugar, R., Schoenfelder, S., Fraser, P., & Luscombe, N. M. (2017). GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. *PLoS One, 12*(4), e0174744. doi: 10.1371/journal.pone.0174744

Mifsud, B., Tavares-Cadete, F., Young, A. N., & Sugar, R. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *47*(6), 598-606. doi: 10.1038/ng.3286

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., . . . Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature, 448*(7153), 553-560. doi: 10.1038/nature06008

Monier, K., Armas, J. C., Etteldorf, S., Ghazal, P., & Sullivan, K. F. (2000). Annexation of the interchromosomal space during viral infection. *Nat Cell Biol, 2*(9), 661-665. doi: 10.1038/35023615

Montag, C., Wagner, J., Gruska, I., & Hagemeier, C. (2006). Human cytomegalovirus blocks tumor necrosis factor alpha- and interleukin-1beta-mediated NF-kappaB signaling. *J Virol, 80*(23), 11686-11698. doi: 10.1128/jvi.01168-06

Moody, C. A., & Laimins, L. A. (2010). Human papillomavirus oncoproteins: pathways to transformation. *Nat Rev Cancer, 10*(8), 550-560. doi: 10.1038/nrc2886

Munger, K., & Jones, D. L. (2015). Human papillomavirus carcinogenesis: an identity crisis in the retinoblastoma tumor suppressor pathway. *J Virol, 89*(9), 4708-4711. doi: 10.1128/jvi.03486-14

Munoz, N., Castellsague, X., de Gonzalez, A. B., & Gissmann, L. (2006). Chapter 1: HPV in the etiology of human cancer. *Vaccine, 24 Suppl 3*, S3/1-10. doi: 10.1016/j.vaccine.2006.05.115

Murphy, J. C., Fischle, W., Verdin, E., & Sinclair, J. H. (2002). Control of cytomegalovirus lytic gene expression by histone acetylation. *Embo j, 21*(5), 1112-1120. doi: 10.1093/emboj/21.5.1112

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., . . . Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature, 502*(7469), 59-64. doi: 10.1038/nature12593

Nagano, T., Varnai, C., Schoenfelder, S., Javierre, B. M., Wingett, S. W., & Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol, 16*, 175. doi: 10.1186/s13059-015-0753-7

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., . . . Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature, 485*(7398), 381-385. doi: 10.1038/nature11049

Ojesina, A. I., Lichtenstein, L., Freeman, S. S., Pedamallu, C. S., Imaz-Rosshandler, I., Pugh, T. J., . . . Meyerson, M. (2014). Landscape of genomic alterations in cervical carcinomas. *Nature, 506*(7488), 371-375. doi: 10.1038/nature12881

Ong, C. T., & Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet, 12*(4), 283-293. doi: 10.1038/nrg2957

Ong, C. T., & Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet, 15*(4), 234-246. doi: 10.1038/nrg3663

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., . . . Fraser, P. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet, 36*(10), 1065-1071. doi: 10.1038/ng1423

Osborne, C. S., Chakalova, L., Mitchell, J. A., Horton, A., Wood, A. L., Bolland, D. J., . . . Fraser, P. (2007). Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol, 5*(8), e192. doi: 10.1371/journal.pbio.0050192

Palstra, R. J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B., & de Laat, W. (2008). Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS One, 3*(2), e1661. doi: 10.1371/journal.pone.0001661

Papantonis, A., Kohro, T., Baboo, S., Larkin, J. D., Deng, B., Short, P., . . . Cook, P. R. (2012). TNFalpha signals through specialized factories where responsive coding and miRNA genes are transcribed. *Embo j, 31*(23), 4404-4414. doi: 10.1038/emboj.2012.288

Parada, L. A., McQueen, P. G., & Misteli, T. (2004). Tissue-specific spatial organization of genomes. *Genome Biol, 5*(7), R44. doi: 10.1186/gb-2004-5-7-r44

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., . . . Merkenschlager, M. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell, 132*(3), 422-433. doi: 10.1016/j.cell.2008.01.011

Paris, C., Pentland, I., Groves, I., Roberts, D. C., Powis, S. J., Coleman, N., . . . Parish, J. L. (2015). CCCTC-binding factor recruitment to the early region of the human papillomavirus 18 genome regulates viral oncogene expression. *J Virol, 89*(9), 4770-4785. doi: 10.1128/jvi.00097-15

Pasquali, L., Gaulton, K. J., Rodriguez-Segui, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., . . . Ferrer, J. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet, 46*(2), 136-143. doi: 10.1038/ng.2870

Paulus, C., & Nevels, M. (2009). The human cytomegalovirus major immediate-early proteins as antagonists of intrinsic and innate antiviral host responses. *Viruses, 1*(3), 760-779. doi: 10.3390/v1030760

Pereira, L., Maidji, E., Fisher, S. J., McDonagh, S., & Tabata, T. (2007). HCMV persistence in the population: potential transplacental transmission. In A. Arvin, G. Campadelli-Fiume, E. Mocarski, P. S. Moore, B. Roizman, R. Whitley & K. Yamanishi (Eds.), *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*. Cambridge: Cambridge University Press

Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., . . . Guilak, F. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *10*(10), 973-976. doi: 10.1038/nmeth.2600

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W., Solovei, I., Brugman, W., . . . van Steensel, B. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell, 38*(4), 603-613. doi: 10.1016/j.molcel.2010.03.016

Peters, J. M., & Nishiyama, T. (2012). Sister chromatid cohesion. *Cold Spring Harb Perspect Biol, 4*(11). doi: 10.1101/cshperspect.a011130

Pett, M., & Coleman, N. (2007). Integration of high-risk human papillomavirus: a key event in cervical carcinogenesis? *J Pathol, 212*(4), 356-367. doi: 10.1002/path.2192

Pett, M. R., Alazawi, W. O., Roberts, I., Dowen, S., Smith, D. I., Stanley, M. A., & Coleman, N. (2004). Acquisition of high-level chromosomal instability is associated with integration of human papillomavirus type 16 in cervical keratinocytes. *Cancer Res, 64*(4), 1359-1368.

Phillips-Cremins, J. E., Sauria, M. E., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S., . . . Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell, 153*(6), 1281-1295. doi: 10.1016/j.cell.2013.04.053

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M., & van Steensel, B. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat Genet, 38*(9), 1005-1014. doi: 10.1038/ng1852

Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., & Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob Health, 4*(9), e609-616. doi: 10.1016/s2214-109x(16)30143-7

Poleshko, A., & Katz, R. A. (2014). Specifying peripheral heterochromatin during nuclear lamina reassembly. *Nucleus, 5*(1), 32-39. doi: 10.4161/nucl.28167

Pomerantz, M. M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M. P., Doddapaneni, H., . . . Freedman, M. L. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet, 41*(8), 882-884. doi: 10.1038/ng.403

Poole, E., & Sinclair, J. (2015). Sleepless latency of human cytomegalovirus. *Med Microbiol Immunol, 204*(3), 421-429. doi: 10.1007/s00430-015-0401-6

Puvion-Dutilleul, F., & Besse, S. (1994). Induction of complete segregation of cellular DNA and non-encapsidated viral genomes in herpes simplex virus type 1 infected HeLa cells as revealed by in situ hybridization. *Chromosoma, 103*(2), 104-110.

Qian, Z., Xuan, B., Gualberto, N., & Yu, D. (2011). The human cytomegalovirus protein pUL38 suppresses endoplasmic reticulum stress-mediated cell death independently of its ability to induce mTORC1 activation. *J Virol, 85*(17), 9103-9113. doi: 10.1128/jvi.00572-11

Raab, J. R., & Kamakaka, R. T. (2010). Insulators and promoters: closer than we think. *Nat Rev Genet, 11*(6), 439-446. doi: 10.1038/nrg2765

Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., . . . Regev, A. (2014). High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell, 159*(7), 1698-1710. doi: 10.1016/j.cell.2014.11.015

Rach, E. A., Winter, D. R., Benjamin, A. M., Corcoran, D. L., Ni, T., Zhu, J., & Ohler, U. (2011). Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet, 7*(1), e1001274. doi: 10.1371/journal.pgen.1001274

Radtke, K., Dohner, K., & Sodeik, B. (2006). Viral interactions with the cytoskeleton: a hitchhiker's guide to the cell. *Cell Microbiol, 8*(3), 387-400. doi: 10.1111/j.1462-5822.2005.00679.x

Ragoczy, T., Bender, M. A., Telling, A., Byron, R., & Groudine, M. (2006). The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. *Genes Dev, 20*(11), 1447-1457. doi: 10.1101/gad.1419506

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell, 159*(7), 1665-1680. doi: 10.1016/j.cell.2014.11.021

Rawlinson, W. D., Farrell, H. E., & Barrell, B. G. (1996). Analysis of the complete DNA sequence of murine cytomegalovirus. *J Virol, 70*(12), 8833-8849.

Reddehase, M. J., Podlech, J., & Grzimek, N. K. (2002). Mouse models of cytomegalovirus latency: overview. *J Clin Virol, 25 Suppl 2*, S23-36.

Reinhardt, B., Winkler, M., Schaarschmidt, P., Pretsch, R., Zhou, S., Vaida, B., . . . Mertens, T. (2006). Human cytomegalovirus-induced reduction of extracellular matrix proteins in vascular smooth muscle cell cultures: a pathomechanism in vasculopathies? *J Gen Virol, 87*(Pt 10), 2849-2858. doi: 10.1099/vir.0.81955-0

Riddell, S. R. (1995). Pathogenesis of cytomegalovirus pneumonia in immunocompromised hosts. *Semin Respir Infect, 10*(4), 199-208.

Roller, R. J., & Baines, J. D. (2017). Herpesvirus Nuclear Egress. *Adv Anat Embryol Cell Biol, 223*, 143-169. doi: 10.1007/978-3-319-53168-7_7

Russell, J., & Zomerdijk, J. C. (2006). The RNA polymerase I transcription machinery. *Biochem Soc Symp*(73), 203-216.

Rutkowski, A. J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., . . . Dolken, L. (2015). Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun, 6*, 7126. doi: 10.1038/ncomms8126

Saksouk, N., Simboeck, E., & Dejardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. *Epigenetics Chromatin, 8*, 3. doi: 10.1186/1756-8935-8-3

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A, 74*(12), 5463-5467.

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature, 489*(7414), 109-113. doi: 10.1038/nature11279

Sarcinella, E., Brown, M., Tellier, R., Petric, M., & Mazzulli, T. (2004). Detection of RNA in purified cytomegalovirus virions. *Virus Res, 104*(2), 129-137. doi: 10.1016/j.virusres.2004.03.008

Satou, Y., Miyazato, P., Ishihara, K., Yaguchi, H., Melamed, A., Miura, M., . . . Bangham, C. R. (2016). The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome. *113*(11), 3054-3059. doi: 10.1073/pnas.1423199113

Scarpini, C. G., Groves, I. J., Pett, M. R., Ward, D., & Coleman, N. (2014). Virus transcript levels and cell growth rates after naturally occurring HPV16 integration events in basal cervical keratinocytes. *J Pathol, 233*(3), 281-293. doi: 10.1002/path.4358

Schardin, M., Cremer, T., Hager, H. D., & Lang, M. (1985). Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Hum Genet, 71*(4), 281-287.

Scheffner, M., Huibregtse, J. M., Vierstra, R. D., & Howley, P. M. (1993). The HPV-16 E6 and E6-AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell, 75*(3), 495-505.

Scheurer, M. E., Tortolero-Luna, G., & Adler-Storthz, K. (2005). Human papillomavirus infection: biology, epidemiology, and prevention. *Int J Gynecol Cancer, 15*(5), 727-746. doi: 10.1111/j.1525-1438.2005.00246.x

Schiller, J. T., Day, P. M., & Kines, R. C. (2010). Current understanding of the mechanism of HPV infection. *Gynecol Oncol, 118*(1 Suppl), S12-17. doi: 10.1016/j.ygyno.2010.04.004

Schmitz, M., Driesch, C., Jansen, L., Runnebaum, I. B., & Durst, M. (2012). Non-random integration of the HPV genome in cervical cancer. *PLoS One, 7*(6), e39632. doi: 10.1371/journal.pone.0039632

Schoenfelder, S., Clay, I., & Fraser, P. (2010a). The transcriptional interactome: gene expression in 3D. *Curr Opin Genet Dev, 20*(2), 127-133. doi: 10.1016/j.gde.2010.02.002

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B. M., . . . Fraser, P. (2015a). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res, 25*(4), 582-597. doi: 10.1101/gr.185272.114

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., . . . Fraser, P. (2010b). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet, 42*(1), 53-61. doi: 10.1038/ng.496

Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B. M., Armstrong, H., Mifsud, B., . . . Elderkin, S. (2015b). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet, 47*(10), 1179-1186. doi: 10.1038/ng.3393

Schones, D. E., Cui, K., Cuddapah, S., Roh, T. Y., Barski, A., Wang, Z., . . . Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell, 132*(5), 887-898. doi: 10.1016/j.cell.2008.02.022

Schubeler, D., MacAlpine, D. M., Scalzo, D., Wirbelauer, C., Kooperberg, C., van Leeuwen, F., . . . Groudine, M. (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev, 18*(11), 1263-1271. doi: 10.1101/gad.1198204

Schubeler, D., Mielke, C., & Bode, J. (1997). Excision of an integrated provirus by the action of FLP recombinase. *In Vitro Cell Dev Biol Anim, 33*(10), 825-830.

Schwalb, B., Michel, M., Zacher, B., Fruhauf, K., Demel, C., Tresch, A., . . . Cramer, P. (2016). TT-seq maps the human transient transcriptome. *Science, 352*(6290), 1225-1228. doi: 10.1126/science.aad9841

Schwammle, V., & Jensen, O. N. (2010). A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. *Bioinformatics, 26*(22), 2841-2848. doi: 10.1093/bioinformatics/btq534

Schwartz, J., & Roizman, B. (1969). Concerning the egress of herpes simplex virus from infected cells: electron and light microscope observations. *Virology, 38*(1), 42-49.

Schwarz, T. M., Volpe, L. A., Abraham, C. G., & Kulesza, C. A. (2013). Molecular investigation of the 7.2 kb RNA of murine cytomegalovirus. *Virol J, 10*, 348. doi: 10.1186/1743-422x-10-348

Scott, R. S. (2017). Epstein-Barr virus: a master epigenetic manipulator. *Curr Opin Virol, 26*, 74-80. doi: 10.1016/j.coviro.2017.07.017

Seedorf, K., Krammer, G., Durst, M., Suhai, S., & Rowekamp, W. G. (1985). Human papillomavirus type 16 DNA sequence. *Virology, 145*(1), 181-185.

Sexton, B. S., Avey, D., Druliner, B. R., Fincher, J. A., Vera, D. L., Grau, D. J., . . . Dennis, J. H. (2014). The spring-loaded genome: nucleosome redistributions are widespread, transient, and DNA-directed. *Genome Res, 24*(2), 251-259. doi: 10.1101/gr.160150.113

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., . . . Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell, 148*(3), 458-472. doi: 10.1016/j.cell.2012.01.010

Shav-Tal, Y., Blechman, J., Darzacq, X., Montagna, C., Dye, B. T., Patton, J. G., . . . Zipori, D. (2005). Dynamic sorting of nuclear components into distinct nucleolar caps during transcriptional inhibition. *Mol Biol Cell, 16*(5), 2395-2413. doi: 10.1091/mbc.E04-11-0992

Shi, X., Hong, T., Walter, K. L., Ewalt, M., Michishita, E., Hung, T., . . . Gozani, O. (2006). ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature, 442*(7098), 96-99. doi: 10.1038/nature04835

Shilatifard, A. (2008). Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation. *Curr Opin Cell Biol, 20*(3), 341-348. doi: 10.1016/j.ceb.2008.03.019

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., . . . Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A, 100*(26), 15776-15781. doi: 10.1073/pnas.2136655100

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet, 15*(4), 272-286. doi: 10.1038/nrg3682

Shulzhenko, N., Lyng, H., Sanson, G. F., & Morgun, A. (2014). Menage a trois: an evolutionary interplay between human papillomavirus, a tumor, and a woman. *Trends Microbiol, 22*(6), 345-353. doi: 10.1016/j.tim.2014.02.009

Simon, M. D., Pinter, S. F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S. K., . . . Lee, J. T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature, 504*(7480), 465-469. doi: 10.1038/nature12719

Simonazzi, G., Curti, A., Cervi, F., Gabrielli, L., Contoli, M., Capretti, M. G., . . . Lazzarotto, T. (2017). Perinatal Outcomes of Non-Primary Maternal Cytomegalovirus Infection: A 15-Year Experience. *Fetal Diagn Ther*. doi: 10.1159/000477168

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., . . . de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet, 38*(11), 1348-1354. doi: 10.1038/ng1896

Sims, R. J., 3rd, Millhouse, S., Chen, C. F., Lewis, B. A., Erdjument-Bromage, H., Tempst, P., . . . Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Mol Cell, 28*(4), 665-676. doi: 10.1016/j.molcel.2007.11.010

Sinclair, J. (2010). Chromatin structure regulates human cytomegalovirus gene expression during latency, reactivation and lytic infection. *Biochim Biophys Acta, 1799*(3-4), 286-295. doi: 10.1016/j.bbagrm.2009.08.001

Sirtori, C., & Bosisio-Bestetti, M. (1967). Nucleolar changes in KB tumor cells infected with herpes simplex virus. *Cancer Res, 27*(2), 367-376.

Sittivarakul, W., & Seepongphun, U. (2017). Incidence Rates and Risk Factors for Vision Loss among AIDS-Related Cytomegalovirus Retinitis Patients in Southern Thailand. *Ocul Immunol Inflamm*, 1-8. doi: 10.1080/09273948.2017.1283044

Skaletskaya, A., Bartle, L. M., Chittenden, T., McCormick, A. L., Mocarski, E. S., & Goldmacher, V. S. (2001). A cytomegalovirus-encoded inhibitor of apoptosis that suppresses caspase-8 activation. *Proc Natl Acad Sci U S A, 98*(14), 7829-7834. doi: 10.1073/pnas.141108798

Smallwood, A., & Ren, B. (2013). Genome organization and long-range regulation of gene expression by enhancers. *Curr Opin Cell Biol, 25*(3), 387-394. doi: 10.1016/j.ceb.2013.02.005

Sofueva, S., Yaffe, E., Chan, W. C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., . . . Hadjur, S. (2013). Cohesin-mediated interactions organize chromosomal domain architecture. *Embo j, 32*(24), 3119-3129. doi: 10.1038/emboj.2013.237

Soufi, A., Garcia, M. F., Jaroszewicz, A., Osman, N., Pellegrini, M., & Zaret, K. S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell, 161*(3), 555-568. doi: 10.1016/j.cell.2015.03.017

Sparmann, A., & van Lohuizen, M. (2006). Polycomb silencers control cell fate, development and cancer. *Nat Rev Cancer, 6*(11), 846-856. doi: 10.1038/nrc1991

Speir, E., Yu, Z. X., Ferrans, V. J., Huang, E. S., & Epstein, S. E. (1998). Aspirin attenuates cytomegalovirus infectivity and gene expression mediated by cyclooxygenase-2 in coronary artery smooth muscle cells. *Circ Res, 83*(2), 210-216.

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet, 13*(9), 613-626. doi: 10.1038/nrg3207

Stanley, M. (2008). Immunobiology of HPV and HPV vaccines. *Gynecol Oncol, 109*(2 Suppl), S15-21. doi: 10.1016/j.ygyno.2008.02.003

Stanley, M., Lowy, D. R., & Frazer, I. (2006). Chapter 12: Prophylactic HPV vaccines: underlying mechanisms. *Vaccine, 24 Suppl 3*, S3/106-113. doi: 10.1016/j.vaccine.2006.05.110

Stanley, M. A., Browne, H. M., Appleby, M., & Minson, A. C. (1989). Properties of a non-tumorigenic human cervical keratinocyte cell line. *Int J Cancer, 43*(4), 672-676.

Stanley, M. A., Pett, M. R., & Coleman, N. (2007). HPV: from infection to cancer. *Biochem Soc Trans, 35*(Pt 6), 1456-1460. doi: 10.1042/bst0351456

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T., Hein, M. Y., Huang, S. X., . . . Weissman, J. S. (2012). Decoding human cytomegalovirus. *Science, 338*(6110), 1088-1093. doi: 10.1126/science.1227919

Stinski, M. F., & Isomura, H. (2008). Role of the cytomegalovirus major immediate early enhancer in acute infection and reactivation from latency. *Med Microbiol Immunol, 197*(2), 223-231. doi: 10.1007/s00430-007-0069-7

Sullivan, A. M., Arsovski, A. A., Lempe, J., Bubb, K. L., Weirauch, M. T., Sabo, P. J., . . . Stamatoyannopoulos, J. A. (2014). Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell Rep, 8*(6), 2015-2030. doi: 10.1016/j.celrep.2014.08.019

Sung, M. H., Baek, S., & Hager, G. L. (2016). Genome-wide footprinting: ready for prime time? *Nat Methods, 13*(3), 222-228. doi: 10.1038/nmeth.3766

Sweet, C. (1999). The pathogenicity of cytomegalovirus. *FEMS Microbiol Rev, 23*(4), 457-482.

Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays, 29*(3), 288-299. doi: 10.1002/bies.20544

Tan, P. Y., Chang, C. W., Chng, K. R., Wansa, K. D., Sung, W. K., & Cheung, E. (2012). Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival. *Mol Cell Biol, 32*(2), 399-414. doi: 10.1128/mcb.05958-11

Tandon, R., & Mocarski, E. S. (2012). Viral and host control of cytomegalovirus maturation. *Trends Microbiol, 20*(8), 392-401. doi: 10.1016/j.tim.2012.04.008

Tang, Q., & Maul, G. G. (2003). Mouse cytomegalovirus immediate-early protein 1 binds with host cell repressors to relieve suppressive effects on viral transcription and replication during lytic infection. *J Virol, 77*(2), 1357-1367.

Tavalai, N., & Stamminger, T. (2011). Intrinsic cellular defense mechanisms targeting human cytomegalovirus. *Virus Res, 157*(2), 128-133. doi: 10.1016/j.virusres.2010.10.002

Terhune, S. S., Schroer, J., & Shenk, T. (2004). RNAs are packaged into human cytomegalovirus virions in proportion to their intracellular concentration. *J Virol, 78*(19), 10390-10398. doi: 10.1128/jvi.78.19.10390-10398.2004

Thomson, I., Gilchrist, S., Bickmore, W. A., & Chubb, J. R. (2004). The radial positioning of chromatin is not inherited through mitosis but is established de novo in early G1. *Curr Biol, 14*(2), 166-172.

Thorland, E. C., Myers, S. L., Gostout, B. S., & Smith, D. I. (2003). Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene, 22*(8), 1225-1237. doi: 10.1038/sj.onc.1206170

Thorland, E. C., Myers, S. L., Persing, D. H., Sarkar, G., McGovern, R. M., Gostout, B. S., & Smith, D. I. (2000). Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites. *Cancer Res, 60*(21), 5916-5921.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., . . . Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature, 489*(7414), 75-82. doi: 10.1038/nature11232

Todaro, G. J., & Green, H. (1963). Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. *J Cell Biol, 17*, 299-313.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics, 25*(9), 1105-1111. doi: 10.1093/bioinformatics/btp120

Ueda, Y., Enomoto, T., Miyatake, T., Ozaki, K., Yoshizaki, T., Kanao, H., . . . Murata, Y. (2003). Monoclonal expansion with integration of high-risk type human papillomaviruses is an initial step for cervical carcinogenesis: association of clonal status and human papillomavirus infection with clinical outcome in cervical intraepithelial neoplasia. *Lab Invest, 83*(10), 1517-1527.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., & Darnell, R. B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science, 302*(5648), 1212-1215. doi: 10.1126/science.1090095

Vakoc, C. R., Mandat, S. A., Olenchock, B. A., & Blobel, G. A. (2005). Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Mol Cell, 19*(3), 381-391. doi: 10.1016/j.molcel.2005.06.011

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., . . . Lander, E. S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*(39). doi: 10.3791/1869

Van Bortle, K., Nichols, M. H., Li, L., Ong, C. T., Takenaka, N., Qin, Z. S., & Corces, V. G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol, 15*(6), R82. doi: 10.1186/gb-2014-15-5-r82

Van Tine, B. A., Kappes, J. C., Banerjee, N. S., Knops, J., Lai, L., Steenbergen, R. D., . . . Chow, L. T. (2004). Clonal selection for transcriptionally active viral oncogenes during progression to cancer. *J Virol, 78*(20), 11172-11186. doi: 10.1128/jvi.78.20.11172-11186.2004

Vastag, L., Koyuncu, E., Grady, S. L., Shenk, T. E., & Rabinowitz, J. D. (2011). Divergent effects of human cytomegalovirus and herpes simplex virus-1 on cellular metabolism. *PLoS Pathog, 7*(7), e1002124. doi: 10.1371/journal.ppat.1002124

Verschure, P. J., van der Kraan, I., de Leeuw, W., van der Vlag, J., Carpenter, A. E., Belmont, A. S., & van Driel, R. (2005). In vivo HP1 targeting causes large-scale chromatin condensation and enhanced histone lysine methylation. *Mol Cell Biol, 25*(11), 4552-4564. doi: 10.1128/mcb.25.11.4552-4564.2005

Vilborg, A., Passarelli, M. C., Yario, T. A., Tycowski, K. T., & Steitz, J. A. (2015). Widespread Inducible Transcription Downstream of Human Genes. *Mol Cell, 59*(3), 449-461. doi: 10.1016/j.molcel.2015.06.016

Vliet-Gregg, P. A., Hamilton, J. R., & Katzenellenbogen, R. A. (2013). NFX1-123 and human papillomavirus 16E6 increase Notch expression in keratinocytes. *J Virol, 87*(24), 13741-13750. doi: 10.1128/jvi.02582-13

Wakamori, M., Fujii, Y., Suka, N., Shirouzu, M., Sakamoto, K., Umehara, T., & Yokoyama, S. (2015). Intra- and inter-nucleosomal interactions of the histone H4 tail revealed with a human

nucleosome core particle with genetically-incorporated H4 tetra-acetylation. *Sci Rep, 5*, 17204. doi: 10.1038/srep17204

Walboomers, J. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V., . . . Munoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol, 189*(1), 12-19. doi: 10.1002/(sici)1096-9896(199909)189:1<12::aid-path431>3.0.co;2-f

Walter, J., Schermelleh, L., Cremer, M., Tashiro, S., & Cremer, T. (2003). Chromosome order in HeLa cells changes during mitosis and early G1, but is stably maintained during subsequent interphase stages. *J Cell Biol, 160*(5), 685-697. doi: 10.1083/jcb.200211103

Wang, & Hayes, J. J. (2008). Acetylation mimics within individual core histone tail domains indicate distinct roles in regulating the stability of higher-order chromatin structure. *Mol Cell Biol, 28*(1), 227-236. doi: 10.1128/mcb.01245-07

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., . . . Fu, X. D. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature, 474*(7351), 390-394. doi: 10.1038/nature10006

Wang, L., Dai, S. Z., Chu, H. J., Cui, H. F., & Xu, X. Y. (2013). Integration sites and genotype distributions of human papillomavirus in cervical intraepithelial neoplasia. *Asian Pac J Cancer Prev, 14*(6), 3837-3841.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., . . . Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature, 420*(6915), 520-562. doi: 10.1038/nature01262

Wathen, M. W., & Stinski, M. F. (1982). Temporal patterns of human cytomegalovirus transcription: mapping the viral RNAs synthesized at immediate early, early, and late times after infection. *J Virol, 41*(2), 462-477.

Watson, Z., Dhummakupt, A., Messer, H., Phelan, D., & Bloom, D. (2013). Role of polycomb proteins in regulating HSV-1 latency. *Viruses, 5*(7), 1740-1757. doi: 10.3390/v5071740

Weekes, M. P., Tomasec, P., Huttlin, E. L., Fielding, C. A., Nusinow, D., Stanton, R. J., . . . Gygi, S. P. (2014). Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell, 157*(6), 1460-1472. doi: 10.1016/j.cell.2014.04.028

Weill, L., Shestakova, E., & Bonnefoy, E. (2003). Transcription factor YY1 binds to the murine beta interferon promoter and regulates its transcriptional capacity with a dual activator/repressor role. *J Virol, 77*(5), 2903-2914.

Weinreb, C., & Raphael, B. J. (2016). Identification of hierarchical chromatin domains. *Bioinformatics, 32*(11), 1601-1609. doi: 10.1093/bioinformatics/btv485

Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., . . . Peters, J. M. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature, 451*(7180), 796-801. doi: 10.1038/nature06634

Wentzensen, N., Vinokurova, S., & von Knebel Doeberitz, M. (2004). Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res, 64*(11), 3878-3884. doi: 10.1158/0008-5472.can-04-0009

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., . . . Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol, 13*(9), R50. doi: 10.1186/gb-2012-13-9-r50

Wiebusch, L., & Hagemeier, C. (1999). Human cytomegalovirus 86-kilodalton IE2 protein blocks cell cycle progression in G(1). *J Virol, 73*(11), 9274-9283.

Wiebusch, L., Neuwirth, A., Grabenhenrich, L., Voigt, S., & Hagemeier, C. (2008). Cell cycle-independent expression of immediate-early gene 3 results in G1 and G2 arrest in murine cytomegalovirus-infected cells. *J Virol, 82*(20), 10188-10198. doi: 10.1128/jvi.01212-08

Winder, D. M., Pett, M. R., Foster, N., Shivji, M. K., Herdman, M. T., Stanley, M. A., . . . Coleman, N. (2007). An increase in DNA double-strand breaks, induced by Ku70 depletion, is associated with human papillomavirus 16 episome loss and de novo viral integration events. *J Pathol, 213*(1), 27-34. doi: 10.1002/path.2206

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., & Andrews, S. (2015). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res, 4*, 1310. doi: 10.12688/f1000research.7334.1

Wright, E., Bain, M., Teague, L., Murphy, J., & Sinclair, J. (2005). Ets-2 repressor factor recruits histone deacetylase to silence human cytomegalovirus immediate-early gene expression in non-permissive cells. *J Gen Virol, 86*(Pt 3), 535-544. doi: 10.1099/vir.0.80352-0

Wutz, G., Varnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R., Tang, W., . . . Peters, J.-M. (2017). CTCF, WAPL and PDS5 proteins control the formation of TADs and loops by cohesin. *bioRxiv*. doi: 10.1101/177444

Xu, J. W., & Ling, S. (2017). Higher-Order Chromatin Regulation of Inflammatory Gene Expression. *2017*, 7848591. doi: 10.1155/2017/7848591

Yamaguchi, Y., Shibata, H., & Handa, H. (2013). Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond. *Biochim Biophys Acta, 1829*(1), 98-104. doi: 10.1016/j.bbagrm.2012.11.007

Ye, R., Su, C., Xu, H., & Zheng, C. (2017). Herpes Simplex Virus 1 Ubiquitin-Specific Protease UL36 Abrogates NF-kappaB Activation in DNA Sensing Signal Pathway. *J Virol, 91*(5). doi: 10.1128/jvi.02417-16

Yim, E. K., & Park, J. S. (2005). The role of HPV E6 and E7 oncoproteins in HPV-associated cervical carcinogenesis. *Cancer Res Treat, 37*(6), 319-324. doi: 10.4143/crt.2005.37.6.319

Yun, W. J., Kim, Y. W., Kang, Y., Lee, J., Dean, A., & Kim, A. (2014). The hematopoietic regulator TAL1 is required for chromatin looping between the beta-globin LCR and human gamma-globin genes to activate transcription. *Nucleic Acids Res, 42*(7), 4283-4293. doi: 10.1093/nar/gku072

Zacapala-Gomez, A. E., Del Moral-Hernandez, O., Villegas-Sepulveda, N., Hidalgo-Miranda, A., Romero-Cordoba, S. L., Beltran-Anaya, F. O., . . . Illades-Aguiar, B. (2016). Changes in global gene expression profiles induced by HPV 16 E6 oncoprotein variants in cervical carcinoma C33-A cells. *Virology, 488*, 187-195. doi: 10.1016/j.virol.2015.11.017

Zalckvar, E., Paulus, C., Tillo, D., Asbach-Nitzsche, A., Lubling, Y., Winterling, C., . . . Nevels, M. (2013). Nucleosome maps of the human cytomegalovirus genome reveal a temporal switch in chromatin organization linked to a major IE protein. *Proc Natl Acad Sci U S A, 110*(32), 13126-13131. doi: 10.1073/pnas.1305548110

Zhan, Y., Mariani, L., Barozzi, I., Schulz, E. G., Bluthgen, N., Stadler, M., . . . Giorgetti, L. (2017). Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res*. doi: 10.1101/gr.212803.116

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol, 9*(9), R137. doi: 10.1186/gb-2008-9-9-r137

Zhao, J. W., Fang, F., Guo, Y., Zhu, T. L., Yu, Y. Y., Kong, F. F., . . . Li, F. (2016). HPV16 integration probably contributes to cervical oncogenesis through interrupting tumor suppressor genes and inducing chromosome instability. *J Exp Clin Cancer Res, 35*(1), 180. doi: 10.1186/s13046-016-0454-4

Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., . . . Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet, 38*(11), 1341-1347. doi: 10.1038/ng1891

Zheng, Z. M., & Baker, C. C. (2006). Papillomavirus genome structure, expression, and post-transcriptional regulation. *Front Biosci, 11*, 2286-2302.

Zhou, Goren, A., & Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet, 12*(1), 7-18. doi: 10.1038/nrg2905

Zhou, & Troyanskaya, O. G. (2016). Probabilistic modelling of chromatin code landscape reveals functional diversity of enhancer-like chromatin states. *Nat Commun, 7*, 10528. doi: 10.1038/ncomms10528

References

Zhou, Z. H., Prasad, B. V., Jakana, J., Rixon, F. J., & Chiu, W. (1994). Protein subunit structures in the herpes simplex virus A-capsid determined from 400 kV spot-scan electron cryomicroscopy. *J Mol Biol, 242*(4), 456-469. doi: 10.1006/jmbi.1994.1594

Ziegert, C., Wentzensen, N., Vinokurova, S., Kisseljov, F., Einenkel, J., Hoeckel, M., & von Knebel Doeberitz, M. (2003). A comprehensive analysis of HPV integration loci in anogenital lesions combining transcript and genome-based amplification techniques. *Oncogene, 22*(25), 3977-3984. doi: 10.1038/sj.onc.1206629

zur Hausen, H. (1991). Viruses in human cancers. *Science, 254*(5035), 1167-1173.